arXiv:1302.2345v1 [math.ST] 10 Feb 2013

Non parametric finite translation mixtures with dependent regime

Elisabeth Gassiat
Laboratoire de Mathématique, Université Paris-Sud and CNRS, Orsay, France,

Judith Rousseau,
ENSAE-CREST and CEREMADE, Université Paris-Dauphine, Paris, France.

February 12, 2013

Abstract

In this paper we consider non parametric finite translation mixtures. We prove that all the parameters of the model are identifiable as soon as the matrix that defines the joint distribution of two consecutive latent variables is non singular and the translation parameters are distinct. Under this assumption, we provide a consistent estimator of the number of populations, of the translation parameters and of the distribution of two consecutive latent variables, which we prove to be asymptotically normally distributed under mild dependency assumptions. We propose a non parametric estimator of the unknown translated density. In case the latent variables form a Markov chain (Hidden Markov models), we prove an oracle inequality leading to the fact that this estimator is minimax adaptive over regularity classes of densities.

Keywords: translation mixtures; non parametric estimation; semi-parametric models; Hidden Markov models,dependent latent variable models.

Short title: Non parametric finite translation mixtures

1 Introduction

Finite mixtures are widely used in applications to model heterogeneous data and to do unsupervised clustering, see for instance MacLachlan and Peel (2000) or Marin et al. (2005) for a review. Latent class models, hidden Markov models or more generally regime switching models may be viewed as mixture models. Finite mixtures are therefore to be understood as convex combinations of a finite number of probability distributions over the space the data lives in, including both static (when the latent variables are independent) and dynamical models. Most of the developed methods use a finite dimensional description of the probability distributions, which requires some prior knowledge of the phenomenon under investigation. In particular applications, it has been noticed that this may lead to poor results and various extensions have been considered. The first natural extension is to consider mixtures with an unknown number of components. This has been extensively studied and used in the literature both from a Bayesian or frequentist point of view, see Akaike (1973), Richardson and Green (1997), Ishwaran et al. (2001), Chambaz and Rousseau (2008), Chambaz et al. (2009), Gassiat and van Handel (pear), to name but a few. However when the emission distribution, i.e. the distribution of each component, is misspecified this results in an overestimation of the number of components, as explained in the discussion in Rabiner (1989). Thus, there has recently been interest in considering nonparametric mixture models in various applications, see for instance the discussion on the Old faithfull dataset in Azzaline and Bowman (1990), the need for nonparametric emission distributions in climate state identification in Lambert et al. (2003) or the nonparametric hidden Markov model proposed in Yau et al. (2011). In absence of training data, mixture models with nonparametric emission distributions are in general not identifiable without additional structural constraints. In a seminal paper, Hall and Zhou (2003) discussed identifiability issues in a 2 -component nonparametric mixture model under repeated measurements (or multivarate) and showed that identifiability essentially only occured if there is at least 3 repeated measurements for each individual. This work has been extended by various authors including Kasahara and Shimotsu (2007), Bonhomme et al. (2011) and references therein. Identifiability recent results about mixtures may also be found in Allman et al. (2009).

Consider location models

$$Y_i = m_{S_i} + \epsilon_i, \quad i \in \mathbb{N} \tag{1.1}$$

where $(S_i)_{i\in\mathbb{N}}$ is an unobserved sequence of random variables with finite state space $\{1,\ldots,k\}$, $(\epsilon_i)_{i\in\mathbb{N}}$ is a sequence of independent identically distributed random variables taking values in $\mathbb{R}$, and $m_j \in \mathbb{R}$, $j = 1,\ldots,k$. The aim is to estimate the parameters k, $m_1,\ldots,m_k$, the distribution of the latent variables $(S_i)_{i\in\mathbb{N}}$ and the distribution F of the ϵ_i's. As usual for finite mixtures, one may recover the parameters only up to relabelling, and obviously, F may only be estimated up to a translation (that would be reversly reported to the m_j's). However the identifiability issue is much more serious without further assumptions. To illustrate the identifiability issues that arise with such models, assume that the S_i's are independent and identically distributed. Then the Y_i's are independent and have distribution

$$P_{\mu,F}(.) = \sum_{j=1}^{k} \mu(j) F\left(\cdot - m_j\right). \tag{1.2}$$

Here, $\mu(j) \geq 0$, $j = 1,\ldots,k$, $\sum_{j=1}^{k}\mu(j) = 1$, $m_j \in \mathbb{R}$, $j = 1,\ldots,k$, and F is a probability distribution on $\mathbb{R}$. An equivalent representation of (1.2) corresponds for instance to $k = 1$, $m_1 = 0$ and $F = P_{\mu,F}$ the marginal distribution. Hunter et al. (2004) have considered model (1.2) with the additional assumption that F is symmetrical and under some constraints on the m_j, in the case of $k \leq 4$, see also L. Bordes and Vandekerkhove (2006) and Butucea and Vandekerkhove (2011) in the case where $k = 2$ for an estimation procedure and asymptotic results.

In this paper, we investigate model (1.1) where the observed variables are not independent and may be non stationary. Interestingly, contrarywise to the independent case, we obtain identifiability without any assumption on F under some very mild conditions on the process $S_1,\cdots,S_n$, see Theorem 2.1. To be precise, if Q is the $k\times k$-matrix such that $Q_{i,j}$ is the probability that $S_1 = i$ and $S_2 = j$, we prove that the knowledge of the distribution of (Y_1,Y_2) allows the identification of k, $m_1,\ldots,m_k$, Q and F as soon as Q is a non singular matrix, whatever F may be. Building upon our identifiability result, we propose an estimator of k, and of the parametric part of the distribution, namely Q and $m_1,\ldots,m_k$. Here, we do not need the sequence $(X_i)_{i\in\mathbb{N}}$ to be strictly stationary and asymptotic stationarity is enough, then Q is the stationary joint disribution of two consecutive latent variables. Moreover, we prove that our estimator is $\sqrt{n}$-consistent, with asymptotic Gaussian distribution, under mild dependency assumptions, see Theorem 3.1. When the number of populations is known and if the translation parameters m_j, $j \leq k$ are known to be bounded by a given constant, we prove that the estimator (centered and at $\sqrt{n}$-scale) has a subgaussian distribution, see Theorem 3.2.

In the context of hidden Markov models as considered in Yau et al. (2011), we propose an estimator of the non parametric part of the distribution, namely F, assuming that it is absolutely continuous with respect to Lebesgue measure. This estimator uses the model selection approach developped in Massart (2007), with the penalized estimated pseudo likelihood contrast based on marginal densities $\sum_{j=1}^{k}\hat{\mu}(j)f(y-\hat{m}_j)$. We prove an oracle inequality, see Theorem 4.1, which allows to deduce that our non parametric estimator is adaptive over regular classes of densities, see Theorem 4.2 and Corollary 1.

The organization of the paper is the following. In section 2 we present and prove our general identifiability theorem. In section 3 we define an estimator of the order and of the parametric part, and state the convergence results: asymptotic gaussian distribution and deviation inequalities. In section 4, we explain our non parametric estimator of the density of F using model selection methods, and state an oracle inequality and adaptive convergence results. Most of the proofs are given in the Appendices.

2 General identifiability result

Let $\mathcal{Q}_k$ be the set of probability mass functions on $\{1,\dots,k\}^2$, that is the set of $k\times k$ matrices $Q=(Q_{i,j})_{1\le i,j\le k}$ such that for all $(i,j)\in\{1,\dots,k\}^2$, $Q_{i,j}\ge 0$, and $\sum_{i=1}^k\sum_{j=1}^k Q_{i,j}=1$. We consider the joint distribution of (Y_1,Y_2) under model (1.1), which has distribution

$$P_{\theta,F}(A\times B)=\sum_{i,j=1}^{k} Q_{ij}F(A-m_i)F(B-m_j),\quad \forall A,B\in\mathcal{B}_{\mathbb{R}} \tag{2.1}$$

where $\mathcal{B}_{\mathbb{R}}$ denotes the Borel σ field of $\mathbb{R}$ and $\theta=\left(m,(Q_{i,j})_{1\le i,j\le k,(i,j)\neq(k,k)}\right)$, with $m=(m_1,\dots,m_k)\in\mathbb{R}^k$. Recall that in this case, ordering the coefficients $m_1\le m_2\le\cdots\le m_k$ and replacing F by $F(.-m_1)$ leads to the same model so that without loss of generality we fixe $0=m_1\le m_2\le\cdots\le m_k$. Let Θ_k be the set of parameters θ such that $m_1=0\le m_2\le\ldots\le m_k$ and $Q\in\mathcal{Q}_k$, where $Q=(Q_{i,j})_{1\le i,j\le k}$, $Q_{k,k}=1-\sum_{(i,j)\neq(k,k)}Q_{i,j}$.
Let also Θ_k^0 be the set of parameters $\theta=\left(m,(Q_{i,j})_{1\le i,j\le k,(i,j)\neq(k,k)}\right)\in\Theta_k$ such that $m_1=0<m_2<\ldots<m_k$ and $\det(\mathrm{Q})\neq 0$. We then have the following result on the identification of F and θ from $P_{\theta,F}$.

Theorem 2.1 *Let F and $\tilde{F}$ be any probability distributions on $\mathbb{R}$. Let k and $\tilde{k}$ be positive integers. If $\theta\in\Theta_k^0$ and $\tilde\theta\in\Theta_{\tilde k}^0$, then*

$$P_{\theta,F}=P_{\tilde\theta,\tilde F}\Longrightarrow k=\tilde k,\ \theta=\tilde\theta \text{ and } \mathrm{F}=\tilde{\mathrm{F}}.$$

Remark 1 *In the same way, it is possible to identify ℓ-marginals, for any $\ell\ge 2$, that is the distribution of $(S_1,\dots,S_\ell)$, m and F on the basis of the distribution of $(Y_1,\dots,Y_\ell)$.*

Remark 2 *The independent case considered in Hunter et al. (2004), L. Bordes and Vandekerkhove (2006), Butucea and Vandekerkhove (2011) is a special case where* $\det(Q)=0$ *for which our identifiability result does not hold. An important class of models is that of hidden Markov models. In that case, if Q is the stationary distribution of two consecutive variables of the hidden Markov chain,* $\det(Q)\neq 0$ *if and only if the transition matrix is non singular and the stationary distribution gives positive weights to each point. When $k=2$, we thus have* $\det(Q)\neq 0$ *if and only if S_1 and S_2 are not independent.*

Proof of Theorem 2.1

Denote by ϕ_F the characteristic function of F, $\phi_{\tilde F}$ the characteristic function of $\tilde F$, $\phi_{\theta,1}$ (respectively $\phi_{\tilde\theta,1}$) the characteristic function of the distribution of m_{S_1} under $P_{\theta,F}$ (respectively under $P_{\tilde\theta,\tilde F}$), $\phi_{\theta,2}$ (respectively $\phi_{\tilde\theta,2}$) the characteristic function of the distribution of m_{S_2} under $P_{\theta,F}$ (respectively under $P_{\tilde\theta,\tilde F}$), and Φ_θ (respectively $\Phi_{\tilde\theta}$) the characteristic function of the distribution of (m_{S_1},m_{S_2}) under $P_{\theta,F}$ (respectively under $P_{\tilde\theta,\tilde F}$). Then since the distribution of Y_1 is the same under $P_{\theta,F}$ and $P_{\tilde\theta,\tilde F}$, one gets that for any $t\in\mathbb{R}$,

$$\phi_F(t)\,\phi_{\theta,1}(t)=\phi_{\tilde F}(t)\,\phi_{\tilde\theta,1}(t)\,. \tag{2.2}$$

Similarly, for any $t\in\mathbb{R}$,

$$\phi_F(t)\,\phi_{\theta,2}(t)=\phi_{\tilde F}(t)\,\phi_{\tilde\theta,2}(t)\,. \tag{2.3}$$

Since the distribution of (Y_1, Y_2) is the same under $P_{\theta,F}$ and $P_{\tilde{\theta},\tilde{F}}$, one gets that for any $\mathbf{t} = (t_1, t_2) \in \mathbb{R}^2$,

$$\phi_F(t_1)\,\phi_F(t_2)\,\Phi_\theta(\mathbf{t}) = \phi_{\tilde{F}}(t_1)\,\phi_{\tilde{F}}(t_2)\,\Phi_{\tilde{\theta}}(\mathbf{t}). \tag{2.4}$$

There exists a neighborhood V of 0 such that for all $t \in V$, $\phi_F(t) \neq 0$, so that (2.2), (2.3) and (2.4) imply that for any $\mathbf{t} = (t_1, t_2) \in V^2$,

$$\Phi_\theta(\mathbf{t})\,\phi_{\tilde{\theta},1}(t_1)\,\phi_{\tilde{\theta},2}(t_2) = \Phi_{\tilde{\theta}}(\mathbf{t})\,\phi_{\theta,1}(t_1)\,\phi_{\theta,2}(t_2). \tag{2.5}$$

Let t_1 be a fixed real number in V. $\Phi_\theta(t_1, t_2)$, $\phi_{\tilde{\theta},2}(t_2)$, $\Phi_{\tilde{\theta}}(t_1, t_2)$, $\phi_{\theta,2}(t_2)$ have analytic continuations for all complex numbers z_2, $\Phi_\theta(t_1, z_2)$, $\phi_{\tilde{\theta}}(z_2)$, $\Phi_{\tilde{\theta}}(t_1, z_2)$, $\phi_\theta(z_2)$ which are entire functions so that (2.5) holds with z_2 in place of t_2 for all z_2 in the complex plane $\mathbb{C}$ and any $t_1 \in V$. Again, let z_2 be a fixed complex number in $\mathbb{C}$. $\Phi_\theta(t_1, z_2)$, $\phi_{\tilde{\theta},1}(t_1)$, $\Phi_{\tilde{\theta}}(t_1, z_2)$, $\phi_{\theta,1}(t_1)$ have analytic continuations $\Phi_\theta(z_1, z_2)$, $\phi_{\tilde{\theta}}(z_1)$, $\Phi_{\tilde{\theta}}(z_1, z_2)$, $\phi_\theta(z_1)$ which are entire functions so that (2.5) holds with z_1 in place of t_1 and z_2 in place of t_2 for all $(z_1, z_2) \in \mathbb{C}^2$. Let now $\mathcal{Z}$ be the set of zeros of $\phi_{\theta,1}$, $\tilde{\mathcal{Z}}$ be the set of zeros of $\phi_{\tilde{\theta},1}$ and fix $z_1 \in \mathcal{Z}$. Then, for any $z_2 \in \mathbb{C}$,

$$\Phi_\theta(z_1, z_2)\,\phi_{\tilde{\theta},1}(z_1)\,\phi_{\tilde{\theta},2}(z_2) = 0. \tag{2.6}$$

We now prove that $z_2 \to \Phi_\theta(z_1, \cdot)$ is not the null function. For any $z \in \mathbb{C}$,

$$\Phi_\theta(z_1, z) = \sum_{\ell=1}^{k} \left[\sum_{j=1}^{k} Q_{\ell,j} e^{i m_j z_1}\right] e^{i m_\ell z}.$$

Since $0 = m_1 < m_2 < \ldots < m_k$, if $\Phi_\theta(z_1, \cdot)$ was the null function, we would have for all $\ell = 1, \ldots, k$

$$\sum_{j=1}^{k} Q_{\ell,j} e^{i m_j z_1} = 0,$$

which is impossible since $\det(Q) \neq 0$. Thus, $\Phi_\theta(z_1, \cdot)$ is an entire function which has isolated zeros, $\phi_{\tilde{\theta},2}(\cdot)$ also, and it is possible to choose z_2 in $\mathbb{C}$ such that $\Phi_\theta(z_1, z_2) \neq 0$ and $\phi_{\tilde{\theta},2}(z_2) \neq 0$. Then (2.6) leads to $\phi_{\tilde{\theta},1}(z_1) = 0$, so that $\mathcal{Z} \subset \tilde{\mathcal{Z}}$. A symmetric argument gives $\tilde{\mathcal{Z}} \subset \mathcal{Z}$ so that $\mathcal{Z} = \tilde{\mathcal{Z}}$. Moreover, $\phi_{\theta,1}$ and $\phi_{\tilde{\theta},1}$ have growth order 1, so that using Hadamard's factorization Theorem (see Stein and Shakarchi (2003) Theorem 5.1) one gets that there exists a polynomial R of degree ≤ 1 such that for all $z \in \mathbb{C}$,

$$\phi_{\theta,1}(z) = e^{R(z)}\phi_{\tilde{\theta},1}(z).$$

But using $\phi_{\theta,1}(0) = \phi_{\tilde{\theta},1}(0) = 1$ we get that there exists a complex number a such that $\phi_{\tilde{\theta},1}(z) = e^{az}\phi_{\theta,1}(z)$. Using now $0 = m_1 < m_2 < \ldots < m_k$, and $0 = \tilde{m}_1 < \tilde{m}_2 < \ldots < \tilde{m}_{\tilde{k}}$ we get that $\phi_{\theta,1} = \phi_{\tilde{\theta},1}$. Similar arguments lead to $\phi_{\theta,2} = \phi_{\tilde{\theta},2}$. Combining this with (2.5) we obtain $\Phi_\theta = \Phi_{\tilde{\theta}}$ which in turns implies that $k = \tilde{k}$ and $\theta = \tilde{\theta}$. Thus, using (2.2), for all $t \in \mathbb{R}$ such that $\phi_{\theta,1}(t) \neq 0$, $\phi_F(t) = \phi_{\tilde{F}}(t)$. Since $\phi_{\theta,1}$ has isolated zeros and ϕ_F, $\phi_{\tilde{F}}$ are continuous functions, one gets $\phi_F = \phi_{\tilde{F}}$ so that $F = \tilde{F}$. □

3 Estimation of the parametric part

3.1 Assumptions on the model

Hereafter, we are given a sequence $(Y_i)_{i\in\mathbb{N}}$ of real random variables with distribution $\mathbb{P}^\star$. We assume that (1.1) holds, with $(S_i)_{i\in\mathbb{N}}$ a sequence of non-observed random variables taking values in $\{1, \ldots, k^\star\}$. We denote by $F^\star$ the common probability distribution of the ϵ_i's, and $m^\star \in \mathbb{R}^{k^\star}$ the possible values of the m_{S_i}'s. We assume:

(A1) (S_i, S_{i+1}) converges in distribution to $Q^\star \in \mathcal{Q}_{k^\star}$.
For $\theta^\star = (m^\star, (Q^\star_{i,j})_{(i,j)\neq(k^\star,k^\star)})$, $\theta^\star \in \Theta^0_{k^\star}$, and all differences $m^\star_j - m^\star_i$, $i, j = 1, \ldots, k^\star$, $i \neq j$, are distinct.

We do not assume that $k^\star$ is known, so that the aim is to estimate $\theta^\star$ and $k^\star$ altogether. Assumption (**A1**) implies that the marginal distributions in $Q^\star$ are identical so that we write from now on $\phi_{\theta^\star} = \phi_{\theta^\star,1} = \phi_{\theta^\star,2}$.

The idea to estimate $\theta^\star$ and k^* is to use equation (2.5) which holds if and only if the parameters are equal. Consider w any probability density on $\mathbb{R}^2$ with compact support $\mathcal{S}$, positive on $\mathcal{S}$ and with 0 belonging to the interior of $\mathcal{S}$; typically $\mathcal{S} = [-a, a]^2$ for some positive a. Define, for any integer k and $\theta \in \Theta_k$:

$$M(\theta) = \int_{\mathbb{R}^2} \left|\Phi_{\theta^\star}(t_1,t_2)\,\phi_{\theta,1}(t_1)\,\phi_{\theta,2}(t_2) - \Phi_\theta(t_1,t_2)\,\phi_{\theta^\star}(t_1)\,\phi_{\theta^\star}(t_2)\right|^2 \\ \left|\phi_{F^\star}(t_1)\,\phi_{F^\star}(t_2)\right|^2 w(t_1,t_2)\,dt_1dt_2. \quad (3.1)$$

We shall use $M(\theta)$ as a contrast function. Indeed, thanks to Theorem 2.1, $\theta \in \Theta_k^0$ is such that $M(\theta) = 0$ if and only if $k = k^\star$ and $\theta = \theta^\star$.
We estimate $M(\cdot)$ by

$$M_n(\theta) = \int_{\mathbb{R}^2} \left|\widehat{\Phi}_n(t_1,t_2)\,\phi_{\theta,1}(t_1)\,\phi_{\theta,2}(t_2) - \Phi_\theta(t_1,t_2)\,\widehat{\phi}_{n,1}(t_1)\,\widehat{\phi}_{n,2}(t_2)\right|^2 w(t_1,t_2)\,dt_1dt_2, \quad (3.2)$$

where $\widehat{\Phi}_n$ is an estimator of the characteristic function of the asymptotic distribution of (Y_t, Y_{t+1}), $\widehat{\phi}_{n,1}(t) = \widehat{\Phi}_n(t,0)$ and $\widehat{\phi}_{n,2}(t) = \widehat{\Phi}_n(0,t)$. One may take for instance the empirical estimator

$$\widehat{\Phi}_n(t_1,t_2) = \frac{1}{n}\sum_{j=1}^{n-1} \exp i\,(t_1Y_j + t_2Y_{j+1}). \quad (3.3)$$

We require that $\widehat{\Phi}_n$ is uniformly upper bounded; if $\widehat{\Phi}_n$ is defined by (3.3) then it is uniformly upper bounded by 1. Define, for any $\mathbf{t} = (t_1,t_2) \in \mathbb{R}^2$

$$Z_n(\mathbf{t}) = \sqrt{n}\left(\widehat{\Phi}_n(\mathbf{t}) - \Phi_{\theta^\star}(\mathbf{t})\phi_{F^\star}(t_1)\,\phi_{F^\star}(t_2)\right).$$

Our main assumptions on the model and on the estimator $\widehat{\Phi}_n$ are the following.

(A2) The process $(Z_n(\mathbf{t}))_{\mathbf{t}\in\mathcal{S}}$ converges weakly to a Gaussian process $(Z(\mathbf{t}))_{\mathbf{t}\in\mathcal{S}}$ in the set of complex continuous functions on $\mathcal{S}$ endowed with the uniform norm and with covariance kernel $\Gamma(\cdot,\cdot)$.

(A3) There exist real numbers E and c (depending on $\theta^\star$) such that for all $x \geq 0$ and $n \geq 1$,

$$\mathbb{P}^\star\left(\sup_{\mathbf{t}\in\mathcal{S}} |Z_n(\mathbf{t})| \geq E + x\right) \leq \exp\left(-cx^2\right).$$

(**A2**) will be used to obtain the asymptotic distribution of the estimator, and (**A3**) to obtain non asymptotic deviation inequalities. Note that (**A2**) and (**A3**) are for instance verified if we use (3.3), under stationarity and mixing conditions on the Y_j's. This follows applying results of Doukhan et al. (1994), Doukhan et al. (1995) and Rio (2000).

3.2 Definition of the estimator

Our contrast function verifies $M(\theta) = 0$ if and only if $\theta = \theta^\star$ only when we restrict θ to belong to $\cup_{k\in\mathbb{N}}\Theta_k^0$. When minimization is performed over $\cup_{k\in\mathbb{N}}\Theta_k^0$ it may happen that the minimizer is on the boundary. To get rid of this problem, we build our estimator $\widehat{\theta}_n$ using a preliminary consistent estimator $\tilde{\theta}_n$, and then restrict the minimization using the information given by $\tilde{\theta}_n$.
Define for any integer k, I_k a positive continuous function on Θ_k^0 and tending to $+\infty$ on the boundary of Θ_k^0 or whenever $\|m\|$ tends to infinity. For instance one may take

$$I_k\left(m, (Q_{i,j})_{(i,j)\neq(k,k)}\right) = -\log \mathrm{detQ} - \sum_{\mathrm{i}=2}^{\mathrm{k}} \log \frac{|\mathrm{m_i} - \mathrm{m_{i-1}}|}{(1 + \|\mathrm{m}\|_\infty)^2}.$$

Let $(k_n, \tilde{\theta}_n)$ be a minimizer over $\{(k,\theta) : k \in \mathbb{N}, \theta \in \Theta_k\}$ of

$$C_n\left(k,\theta\right) = M_n\left(\theta\right) + \lambda_n\left[J\left(k\right) + I_k\left(\theta\right)\right]$$

where $J : \mathbb{N} \to \mathbb{N}$ is an increasing function tending to infinity at infinity and $(\lambda_n)_{n\in\mathbb{N}}$ a decreasing sequence of real numbers tending to 0 at infinity such that

$$\lim_{n\to+\infty} \sqrt{n}\lambda_n = +\infty \tag{3.4}$$

Define now $\widehat{\theta}_n$ as a minimizer of M_n over

$$\left\{\theta \in \Theta_{k_n} \ : \ I_{k_n}\left(\theta\right) \le 2I_{k_n}\left(\tilde{\theta}_n\right)\right\}.$$

In case $k^\star$ is known, we may choose another estimator. Let $\mathcal{K}$ be a compact subset of $\Theta^0_{k^\star}$. We denote by $\overline{\theta}_n(\mathcal{K})$ a minimizer of M_n over $\mathcal{K}$. This estimator will also be used as a theoretical trick in the proof of the asymptotic distribution of $\widehat{\theta}_n$.

3.3 Asymptotic results

Our first result gives the asymptotic distribution of $\widehat{\theta}_n$. To define the asymptotic variance, we define $\nabla M\left(\theta\right)$ the gradient of M at point θ and $D_2M\left(\theta\right)$ the Hessian of M at point θ. We also set V the variance of the gaussian process

$$\int \{C\left(\mathbf{t}\right)\left[Z\left(-\mathbf{t}\right)\phi_{\theta^\star}\left(-t_1\right)\phi_{\theta^\star}\left(-t_2\right) - \Phi_{\theta^\star}\left(-\mathbf{t}\right)\left(Z(-t_1,0)\phi_{\theta^\star}\left(-t_2\right) + Z(0,-t_2)\phi_{\theta^\star}\left(-t_1\right)\right)\right]$$
$$+C\left(-\mathbf{t}\right)\left[Z\left(\mathbf{t}\right)\phi_{\theta^\star}\left(t_1\right)\phi_{\theta^\star}\left(t_2\right) - \Phi_{\theta^\star}\left(\mathbf{t}\right)\left(Z(t_1,0)\phi_{\theta^\star}\left(t_2\right) + Z(0,t_2)\phi_{\theta^\star}\left(t_1\right)\right)\right]\} w\left(\mathbf{t}\right)d\mathbf{t}$$

where

$$C\left(\mathbf{t}\right) = \Phi_{\theta^\star}\left(\mathbf{t}\right)\nabla\left(\phi_{\theta^\star}\left(t_1\right)\phi_{\theta^\star}\left(t_2\right)\right) - \nabla\Phi_{\theta^\star}\left(\mathbf{t}\right)\phi_{\theta^\star}\left(t_1\right)\phi_{\theta^\star}\left(t_2\right).$$

Theorem 3.1 *Assume* (**A1**), (**A2**), *and (3.4). Then $D_2M\left(\theta^\star\right)$ is non singular, and for any compact subset $\mathcal{K}$ of $\Theta^0_{k^\star}$ such that $\theta^\star$ lies in the interior of $\mathcal{K}$, $\sqrt{n}(\overline{\theta}_n(\mathcal{K}) - \theta^*)$ converges in distribution to the centered Gaussian with variance*

$$\Sigma = D_2M\left(\theta^\star\right)^{-1} V D_2M\left(\theta^\star\right)^{-1}.$$

Moreover, $\sqrt{n}(\widehat{\theta}_n - \theta^*)$ converges in distribution to the centered Gaussian with variance Σ.

If one wants to use Theorem 3.1 to build confidence sets, one needs to have a consistent estimator of Σ. Since D_2M is a continuous functions of θ, $D_2M\left(\widehat{\theta}_n\right)$ is a consistent estimator of $D_2M\left(\theta^\star\right)$. Also, V may be viewed as a continuous function of $\Gamma(\cdot,\cdot)$ and θ, as easy but tedious computations show. One may use empirical estimators of $\Gamma(\cdot,\cdot)$ which are uniformly consistent under stationarity and mixing conditions, to get a consistent estimator of V. This leads to a plug-in consistent estimator of Σ.
Another possible way to estimate Σ is to use a boostrap method, following for instance Clemencon et al. (2009) when the hidden variables form a Markov chain.

When we have deviation inequalities for the process Z_n, we are able to provide deviation inequalities for $\sqrt{n}(\overline{\theta}_n(\mathcal{K}) - \theta^*)$. Such inequalities have interest by themselves, they will also be used for proving adaptivity of our non parametric estimator in Section 4.

Theorem 3.2 *Assume* (**A1**) *and* (**A3**). *Let $\mathcal{K}$ be a compact subset of $\Theta^0_{k^\star}$ such that $\theta^\star$ lies in the interior of $\mathcal{K}$. Then there exist real numbers $c^\star$, $M^\star$, and an integer $n^\star$ such that for all $n \ge n^\star$ and $M \ge M^\star$,*

$$\mathbb{P}^\star\left(\sqrt{n}\|\overline{\theta}_n(\mathcal{K}) - \theta^\star\| \ge M\right) \le 8\exp\left(-c^\star M^2\right).$$

In particular, for any integer p,

$$\sup_{n\ge1} E_{\mathbb{P}^\star}\left[\left(\sqrt{n}\|\overline{\theta}_n(\mathcal{K}) - \theta^\star\|\right)^p\right] < +\infty.$$

4 Estimation of the non parametric part in the case of hidden Markov models

In this section we assume that $\mathbb{P}^\star$ is the distribution of a stationary ergodic hidden Markov model (HMM for short), that is the sequence $(S_t)_{t\in\mathbb{N}}$ is a stationary ergodic Markov chain. We also assume that the unknown distribution $F^\star$ has density $f^\star$ with respect to Lebesgue measure. Thus the density $s^\star$ of Y_1 writes

$$s^\star(y) = \sum_{j=1}^{k^\star} \mu^\star(j) f^\star\left(y - m_j^\star\right),$$

where $\mu^\star(j) = \sum_{i=1}^{k^\star} Q_{j,i}^\star$, $1 \leq i \leq k^\star$. We shall assume moreover:

(A4) For all $i, j = 1, \ldots, k^\star$, $Q_{i,j}^\star > 0$, and there exists $\delta > 0$ such that

$$\int_{\mathbb{R}} \left[f^\star(y)\right]^{1-\delta} dy < +\infty.$$

Notice that, if the observations form a stationary HMM and if for all $i, j = 1, \ldots, k^\star$, $Q_{i,j}^\star > 0$, then the sequence is geometrically uniformly ergodic, and applying results of Doukhan et al. (1994), Doukhan et al. (1995) and Rio (2000), (**A2**) and (**A3**) hold if we use (3.3).

We propose to use model selection methods to estimate $f^\star$ using penalized marginal likelihood. We assume in this section that $k^\star$ is known, and that we are given an estimator $\widehat{\theta}_n = ((\widehat{m}_i)_{1\leq i\leq k^\star}, (\widehat{Q}_{i,j})_{(i,j)\neq(k^\star,k^\star)}) = \overline{\theta}_n(\mathcal{K})$ of $\theta^\star$ for some compact subset $\mathcal{K}$ of $\Theta_{k^\star}^0$ such that $\theta^\star$ lies in the interior of $\mathcal{K}$. Let $\widehat{\mu}(i) = \sum_{j=1}^{k^\star} \widehat{Q}_{i,j}$, $1 \leq i \leq k^\star$. Define for any density function f on $\mathbb{R}$

$$\ell_n(f) = \frac{1}{n} \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k^\star} \widehat{\mu}(j) f(Y_i - \widehat{m}_j)\right].$$

Let $\mathcal{F}$ be the set of probability densities on $\mathbb{R}$. We shall use the model collection $(\mathcal{F}_p)_{p\geq 2}$ of Gaussian mixtures with p components as approximation of $\mathcal{F}$. Let us define for any integer p

$$\mathcal{F}_p = \left\{\sum_{i=1}^{p} \pi_i \varphi_{u_i}(x - \alpha_i),\ \alpha_i \in [-A_p, A_p],\ u_i \in [b_p, B],\ \pi_i \geq 0,\ i = 1, \ldots, p,\ \sum_{i=1}^{p} \pi_i = 1\right\} \tag{4.1}$$

where B and A_p, b_p, $p \geq 2$, are positive real numbers, and where φ_β is the Gaussian density with variance β^2 given by $\varphi_\beta(x) = \exp(-x^2/2\beta^2)/\beta\sqrt{2\pi}$. For any $p \geq 2$, let $\widehat{f}_p$ be the maximizer of $\ell_n(f)$ over $\mathcal{F}_p$. Define

$$D_n(p) = -\ell_n\left(\widehat{f}_p\right) + \text{pen}(p, n).$$

Our model selection estimator $\widehat{f}$ will be given by $\widehat{f}_{\widehat{p}}$ whenever $\widehat{p}$ is a minimizer of D_n.

4.1 Oracle inequality

The following theorem says that a suitable choice of the penalty term pen(p, n) leads to an estimator having good non asymptotic and asymptotic properties. In the following,

$$\widehat{s}_{\widehat{p}}(\cdot) = \sum_{j=1}^{k^\star} \widehat{\mu}(j) \widehat{f}_{\widehat{p}}(\cdot - \widehat{m}_j),$$

is the estimator of $s^\star$,

$$S_p^\star = \{\sum_{j=1}^{k^\star} \mu^\star(j) f\left(\cdot - m_j^\star\right), f \in \mathcal{F}_p\}$$

for any $p \geq 2$, $h^2(\cdot,\cdot)$ is the Hellinger distance and $K(\cdot,\cdot)$ the Kullback-Leibler divergence between probability densities. For any $p \geq 1$, fix some $f_p \in \mathcal{F}_p$ and set $s_p = \sum_{j=1}^{k^\star} \mu^\star(j) f_p\left(\cdot - m_j^\star\right)$. Of course to derive good behaviour of the estimator from the oracle inequality, one will have to choose carefully f_p.

Theorem 4.1 *Assume* (**A1**), (**A3**) *and* (**A4**). *Let $(x_p)_{p\geq 2}$ be a sequence of positive real numbers such that $\Sigma = \sum_{p\geq 2} e^{-x_p} < +\infty$. Then there exist positive real numbers κ and C, depending only on $Q^\star$ and δ such that, as soon as*

$$pen(p,n) \geq \frac{\kappa}{n}\left(k^\star p\left[\log n + \log\left(\frac{1}{b_p}\right) + \log A_p\right] + x_p\left[1 + \left|\log\left(1 + \frac{1}{b_p^\delta}\right)\right|\right]\right),$$

one has

$$E_{\mathbb{P}^\star}\left[h^2\left(s^\star, \widehat{s}_{\widehat{p}}\right)\right] \leq C\left\{\inf_{p\geq 2}\left(K\left(f^\star, f_p\right) + pen(p,n) + E_{\mathbb{P}^\star}\left[V_p\right]\right) + \frac{\Sigma}{n}\right\}$$

with

$$V_p = \frac{1}{n}\sum_{i=1}^n \log\left(\frac{\sum_{j=1}^{k^\star} \widehat{\mu}(j) f_p(Y_i - \widehat{m}_j)}{\sum_{j=1}^{k^\star} \mu^\star(j) f_p(Y_i - m_j^\star)}\right).$$

The proof of Theorem 4.1 is postponed to Appendix C.
Notice that the constant in the so-called oracle inequality depends on $\mathbb{P}^\star$, so that the result of Theorem 4.1 is not of real practical use. Also, the upper bound depends on $\widehat{\theta}$, for which the results in Section 3 are for large enough n. However, Theorem 4.1 is the building stone to understand how to choose a penalty function and to prove adaptivity of our estimator.

4.2 Adaptive estimation

We prove now that $\widehat{s}_{\widehat{p}}$ is an adaptive estimator of $s^\star$, and that, if $\max_j \mu^\star(j) > \frac{1}{2}$, $\widehat{f}_{\widehat{p}}$ is an adaptive estimator of $f^\star$. Adaptivity will be proved on the following classes of regular densities.

Let $y_0 > 0$, $c > 0$, $M > 0$, $\tau > 0$, $C > 0$, $\lambda > 0$ and L a positive polynomial function on $\mathbb{R}$. Let also $\beta > 0$ and $\gamma > (3/2 - \beta)_+$. If we denote $\mathcal{P} = (y_0, c_0, M, \tau, C, \lambda, L)$, we define $\mathcal{H}_{loc}(\beta, \gamma, \mathcal{P})$ as the set of probability densities f on $\mathbb{R}$ satisfying:

- f is monotone on $(-\infty, -y_0)$ and on $(y_0, +\infty)$, and $\inf_{|y|\leq y_0} f(y) \geq c_0 > 0$.
- $$\forall y \in \mathbb{R},\ f(y) \leq M e^{-\tau|y|} \tag{4.2}$$
- $\log f$ is $\lfloor\beta\rfloor$ times continuously differentiable with derivatives ℓ_j, $j \leq \beta$ satisfying for all $x \in \mathbb{R}$ and all $|y - x| \leq \lambda$,

  $$|\ell_{\lfloor\beta\rfloor}(y) - \ell_{\lfloor\beta\rfloor}(x)| \leq \lfloor\beta\rfloor! L(x)|y - x|^{\beta - \lfloor\beta\rfloor}$$

  and
  $$\int_{\mathbb{R}} |\ell_j(y)|^{\frac{2\beta+\gamma}{j}} f(y) dy \leq C.$$

We use $\widehat{s}_{\widehat{p}}$ where the penalty is set to

$$\mathrm{pen}(p,n) = \frac{3\kappa}{n}(k^\star p + x_p)\log n.$$

Theorem 4.2 *Assume* (**A1**)*,* (**A3**) *and* (**A4**)*. Then for any* $\mathcal{P}$*,* $\beta \geq 1/2$ *and* $\gamma > (3/2 - \beta)_+$*, there exists* $C(\beta,\gamma,\mathcal{P}) > 0$ *such that*

$$\limsup_{n\to+\infty} \left(\frac{n}{(\log n)^3}\right)^{\frac{2\beta}{2\beta+1}} \sup_{f^\star\in\mathcal{H}_{loc}(\beta,\gamma,\mathcal{P})} E_{\mathbb{P}^\star}\left[h^2\left(s^\star, \widehat{s}_{\widehat{p}}\right)\right] \leq C(\beta,\gamma,\mathcal{P}).$$

Thus, $\widehat{s}_{\widehat{p}}$ is adaptive on the regularity β of the density classes up to $(\log n)^{3\beta/(2\beta+1)}$, see Maugis-Rabusseau and Michel (2012) for a lower bound of the asymptotic minimax risk in the case of independent and identically distributed random variables. Using Theorem 4.2, we can also derive adaptive asymptotic rates for the minimax L_1-risk for the estimation of f^*.

Corollary 1 *Assume* (**A1**)*,* (**A3**)*,* (**A4**) *and that* $\max_j \mu^\star(j) > \frac{1}{2}$ *. Then for any* $\mathcal{P}$*,* $\beta \geq 1/2$ *and* $\gamma > (3/2-\beta)_+$*,*

$$\limsup_{n\to+\infty} \left(\frac{n}{(\log n)^3}\right)^{\frac{\beta}{2\beta+1}} \sup_{f^\star\in\mathcal{H}_{loc}(\beta,\gamma,\mathcal{P})} E_{\mathbb{P}^\star}\left[\left\|\widehat{f}_{\widehat{p}} - f^\star\right\|_1\right] \leq \frac{2\sqrt{C(\beta,\gamma,\mathcal{P})}}{(2\max_j \mu^\star(j) - 1)}.$$

It is possible that the constraint, $\max_j \mu^\star(j) > 1/2$ is not sharp, however note that the Fourier transform of $s^\star$ is expressed as $\phi_{\theta^\star}\phi_{f^\star}$ with $\phi_{\theta^\star}(t) = \sum_{j=1}^k \mu^\star(j)e^{itm_j^\star}$ and $\phi_{f^\star}$ the Fourier transform of $f^\star$, and that $|\phi_{\theta^\star}(t)| > 0$ for all $t \in \mathbb{R}$ if and only if $\max_j \mu^\star(j) > 1/2$, applying the main theorem of Moreno (1973).

Proof of Corollary 1

We shall use

$$\|s^\star - \widehat{s}_{\widehat{p}}\|_1 \leq 2h\left(s^\star, \widehat{s}_{\widehat{p}}\right),$$

together with

$$\begin{aligned}
\|s^\star - \widehat{s}_{\widehat{p}}\|_1 &= \|\sum_{j=1}^{k^\star} \mu^\star(j) f^\star\left(\cdot - m_j^\star\right) - \sum_{j=1}^{k^\star} \widehat{\mu}(j) \widehat{f}_{\widehat{p}}\left(\cdot - \widehat{m}_j\right)\|_1 \\
&\geq \|\sum_{j=1}^{k^\star} \mu^\star(j)\left(\widehat{f}_{\widehat{p}} - f^\star\right)\left(\cdot - \widehat{m}_j\right)\|_1 - \|\widehat{\theta}_n - \theta^\star\| \\
&\quad - \|\sum_{j=1}^{k^\star} \mu^\star(j)\left(f^\star\left(\cdot - m_j^\star\right) - f^\star\left(\cdot - \widehat{m}_j\right)\right)\|_1 \\
&\geq \left(2\max_j \mu^\star(j) - 1\right)\left\|\widehat{f}_{\widehat{p}} - f^\star\right\|_1 - \|\widehat{\theta}_n - \theta^\star\| \\
&\quad - \|f^\star\left(\cdot - m_j^\star\right) - f^\star\left(\cdot - \widehat{m}_j\right)\|_1
\end{aligned}$$

which follows by using iteratively the triangle inequality. Using $\beta \geq 1/2$, Theorem 3.2. and Theorem 4.2, we thus get that

$$\limsup_{n\to+\infty} \left(\frac{n}{(\log n)^3}\right)^{\frac{\beta}{2\beta+1}} \sup_{f^\star\in\mathcal{H}_{loc}(\beta,\gamma,\mathcal{P})} E_{\mathbb{P}^\star}\left[\left\|\widehat{f}_{\widehat{p}} - f^\star\right\|_1\right] \leq \frac{2\sqrt{C(\beta,\gamma,\mathcal{P})}}{(2\max_j \mu^\star(j) - 1)}$$

as soon as

$$\lim_{n\to+\infty} \left(\frac{n}{(\log n)^3}\right)^{\frac{\beta}{2\beta+1}} \sup_{f^\star\in\mathcal{H}_{loc}(\beta,\gamma,\mathcal{P})} E_{\mathbb{P}^\star}\left[\|f^\star\left(\cdot - m_j^\star\right) - f^\star\left(\cdot - \widehat{m}_j\right)\|_1\right] = 0. \tag{4.3}$$

Now, since $f^\star \in H_{loc}(\beta,\gamma,\mathcal{P})$ with $\beta \geq 1/2$, if $|\widehat{m}_j - m_j^\star| \leq \lambda$,

$$|\log f^\star(y - \widehat{m}_j) - \log f^\star(y - m_j^\star)| \leq L(y - m_j^\star)|\widehat{m}_j - m_j^\star|^{\beta\wedge 1}.$$

Set $M \geq \frac{1}{2c^\star}$, and $a > 0$ such that, if $|y| \leq n^a$, then $L(y)|\widehat{m}_j - m_j^\star|^{\beta\wedge 1} \leq 1$. Observe also that since $\widehat{\theta}_n$ stays in a compact set, for large enough n, if $|y| \geq n^a$, then for any j, $|y - \widehat{m}_j| \geq n^a/2$ and $|y - m_j^\star| \geq n^a/2$. We obtain, using $|e^u - 1| \leq 2u$ for $0 \leq u \leq 1$:

$$\begin{aligned} \left\| f^\star\left(\cdot - m_j^\star\right) - f^\star\left(\cdot - \widehat{m}_j\right) \right\|_1 \quad \leq \quad & 2\left(\frac{M\log n}{n}\right)^{-(\beta\wedge 1)/2} \int L(y - m_j^\star) f^\star(y - m_j^\star) dy \\ & + 2\int_{|y|\geq n^a/2} f^\star(y) dy + \mathbb{1}_{\|\theta^\star - \widehat{\theta}_n\| > \sqrt{M\log n}/\sqrt{n}}, \end{aligned}$$

and (4.3) follows from Theorem 3.2, $\beta \geq 1/2$ and the fact that $f^\star \in H_{loc}(\beta, \gamma, \mathcal{P})$ has exponentially decreasing tails. □

4.3 Computation of $\widehat{f_p}$

The computation of $\widehat{f_p}$ may be performed using the EM-algorithm, which is particularly simple for Gaussian mixtures. Indeed, for $f = \sum_{i=1}^{p} \pi_i \varphi_{\beta_i}\left(\cdot - \alpha_i\right)$, $\sum_{j=1}^{k^\star} \widehat{\mu}\left(j\right) f\left(\cdot - \widehat{m}_j\right)$ is a mixture of $pk^\star$ Gaussian densities $\phi_{\beta_i}\left(\cdot - \alpha_i - \widehat{m}_j\right)$ with weights $\pi_i \widehat{\mu}\left(j\right)$. Starting from an initial point $((\pi_i^0)_{1\leq i\leq p}, (v_i^0)_{1\leq i\leq p}, (\sigma_i^0)_{1\leq i\leq p})$, the EM l-th iteration may be easily computed as

$$\pi_i^{l+1} = \frac{\sum_{j=1}^{k^\star}\sum_{t=1}^{n} \widehat{\mu}\left(j\right) \pi_i^l \varphi_{\beta_i^l}\left(Y_t - \widehat{m}_j - \alpha_i^l\right)}{\sum_{i'=1}^{p}\sum_{j=1}^{k^\star}\sum_{t=1}^{n} \widehat{\mu}\left(j\right) \pi_{i'}^l \varphi_{\beta_{i'}^l}\left(Y_t - \widehat{m}_j - \alpha_{i'}^l\right)},\ i = 1, \ldots, p,$$

$$\alpha_i^{l+1} = T_{-A_p, A_p}\left[\frac{\sum_{j=1}^{k^\star}\sum_{t=1}^{n} \left(Y_t - \widehat{m}_j\right) \widehat{\mu}\left(j\right) \pi_i^l \varphi_{\beta_i^l}\left(Y_t - \widehat{m}_j - \alpha_i^l\right)}{\sum_{j=1}^{k^\star}\sum_{t=1}^{n} \widehat{\mu}\left(j\right) \pi_i^l \varphi_{\beta_i^l}\left(Y_t - \widehat{m}_j - \alpha_i^l\right)}\right],\ i = 1, \ldots, p,$$

where for any real numbers C_1, C_2, T_{C_1, C_2} is the troncature function: $T_{C_1,C_2}(x) = x\mathbb{1}_{C_1\leq x\leq C_2} + C_1\mathbb{1}_{x<C_1} + C_2\mathbb{1}_{x>C_2}$, and

$$\sigma_i^{l+1} = T_{b_p, B}\left[\frac{\sum_{j=1}^{k^\star}\sum_{t=1}^{n} \left(Y_t - \widehat{m}_j - v_i^l\right)^2 \widehat{\mu}\left(j\right) \pi_i^l \varphi_{\beta_i^l}\left(Y_t - \widehat{m}_j - \alpha_i^l\right)}{\sum_{j=1}^{k^\star}\sum_{t=1}^{n} \widehat{\mu}\left(j\right) \pi_i^l \varphi_{\beta_i^l}\left(Y_t - \widehat{m}_j - \alpha_i^l\right)}\right],\ i = 1, \ldots, p.$$

Acknowledgements

This work was partly supported by the 2010–2014 grant ANR Banhdits AAP Blanc SIMI 1.

A Proof of Theorem 3.1

First of all, we prove a lemma we shall use several times. Using $||A|^2 - |B|^2| \leq |A - B|||A| + |B||$ and the fact that characteristic functions are uniformly upper bounded by 1, we get that for any integer k and any $\theta \in \Theta_k$:

$$\begin{aligned} \left|M_n\left(\theta\right) - M\left(\theta\right)\right| \leq 2\int \Big\{ & \left|\widehat{\Phi}_n\left(t_1, t_2\right) - \Phi_{\theta^\star}\left(t_1, t_2\right) \phi_{F^\star}\left(t_1\right) \phi_{F^\star}\left(t_2\right)\right| \\ & + \left|\widehat{\phi}_n\left(t_1\right) \widehat{\phi}_n\left(t_2\right) - \phi_{\theta^\star}\left(t_1\right) \phi_{\theta^\star}\left(t_2\right) \phi_{F^\star}\left(t_1\right) \phi_{F^\star}\left(t_2\right)\right| \Big\} w\left(t_1, t_2\right) dt_1 dt_2. \end{aligned}$$

The upper bound does not depend on k and θ, $\widehat{\Phi}_n$ is uniformly upper bounded, and we get

$$\sup_{k\geq 2,\ \theta\in\Theta_k} \left|M_n\left(\theta\right) - M\left(\theta\right)\right| = O\left(\sup_{\mathbf{t}\in\mathcal{S}} \left|\frac{Z_n(\mathbf{t})}{\sqrt{n}}\right|\right) = O_{\mathbb{P}^\star}(1/\sqrt{n}) \qquad \text{(A.1)}$$

which together with Theorem 2.1 gives

Lemma 1 *If $(k_n, \theta_n)_n$, $\theta_n \in \Theta_{k_n}$, is a random sequence such that there exists an integer $K \geq k^*$, and a compact subset $\mathcal{T}$ of $\cup_{k \leq K} \Theta_k^0$ such that*

$$\mathbb{P}^* \left(k_n \leq K \text{ and } \theta_\mathrm{n} \in \mathcal{T}\right) \to 1 \text{ and } \mathrm{M_n}\left(\theta_\mathrm{n}\right) = \mathrm{o}_{\mathbb{P}^\star}(1),$$

then

$$\mathbb{P}^* \left(k_n = k^\star\right) \to 1 \text{ and } \theta_\mathrm{n} = \theta^\star + \mathrm{o}_{\mathbb{P}^\star}(1).$$

Since $C_n\left(k_n, \tilde{\theta}_n\right) \leq C_n\left(k^*, \theta^*\right)$ and M_n is a non negative function, we get

$$\left[J(k_n) + I_{k_n}\left(\tilde{\theta}_n\right)\right] \leq \left[J(k^\star) + I_{k^\star}\left(\theta^\star\right)\right] + \frac{M_n\left(\theta^\star\right) - M\left(\theta^\star\right)}{\lambda_n},$$

so that using (A.1), assumption (**A2**) and (3.4) we get

$$\left[J(k_n) + I_{k_n}\left(\tilde{\theta}_n\right)\right] \leq \left[J(k^\star) + I_{k^\star}\left(\theta^\star\right)\right] + o_{\mathbb{P}^\star}\left(1\right). \tag{A.2}$$

Also,

$$M_n\left(\tilde{\theta}_n\right) \leq M_n\left(\theta^\star\right) + \lambda_n\left[J(k^\star) + I_{k^\star}\left(\theta^\star\right)\right],$$

so that

$$M_n\left(\tilde{\theta}_n\right) = o_{\mathbb{P}^\star}\left(1\right).$$

Thus, using (A.2) and Lemma 1

$$\mathbb{P}^* \left(k_n = k^\star\right) \to 1 \text{ and } \tilde{\theta}_\mathrm{n} = \theta^\star + \mathrm{o}_{\mathbb{P}^\star}(1). \tag{A.3}$$

Set now $\mathcal{K} = \{\theta \in \Theta_{k^\star} : I_{k^\star}\left(\theta\right) \leq 4 I_{k^\star}\left(\theta^\star\right)\}$. $\mathcal{K}$ is a compact subset of $\Theta^0_{k^\star}$. Let E_n be the event ($k_n = k^\star$ and $\widehat{\theta}_n = \overline{\theta}_n(\mathcal{K})$). Using Lemma 1, we get that $\overline{\theta}_n(\mathcal{K})$ is a consistent estimator of $\theta^\star$, and using (A.3) and Lemma 1, we get also that $\widehat{\theta}_n$ is a consistent estimator of $\theta^\star$, so M_n has the same minimizer on $\mathcal{K}$ and on $\{I_{k_n}(\theta) \leq 2 I_{k_n}(\tilde{\theta}_n)\}$, with probability tending to 1, since it belongs to a neigbourhood of θ^*. Thus, $\mathbb{P}^\star\left(E_n\right) \to 1$. Now, since

$$\widehat{\theta}_n = \overline{\theta}_n(\mathcal{K}) 1\!\!1_{E_n} + \widehat{\theta}_n 1\!\!1_{E_n^c},$$

Theorem 3.1 follows as soon as we prove that $\sqrt{n}(\overline{\theta}_n(\mathcal{K}) - \theta^\star)$ converges in distribution to the centered Gaussian with variance Σ. But this is a straighforward consequence of

$$D_2 M_n\left(\theta_n\right)\left(\overline{\theta}_n(\mathcal{K}) - \theta^\star\right) = \nabla M_n\left(\theta^\star\right),$$

for some $\theta_n \in \Theta_{k^\star}$ such that $\|\theta_n - \theta^\star\| \leq \|\overline{\theta}_n(\mathcal{K}) - \theta^\star\|$, the consistency of $\overline{\theta}_n(\mathcal{K})$ and the following Lemma

Lemma 2 *Assume* **(A1)** *and* **(A2)**. *Then*

- *$\sqrt{n} \nabla M_n\left(\theta^\star\right)$ converges in distribution to a centered gaussian with variance V.*
- *$D_2 M\left(\theta^\star\right)$ is non singular, and for any random variable $\theta_n \in \Theta_{k^\star}$ converging in $\mathbb{P}^\star$-probability to $\theta^\star$, one has*

$$D_2 M_n\left(\theta_n\right) = D_2 M\left(\theta^\star\right) + o_{\mathbb{P}^\star}\left(1\right).$$

Proof of Lemma 2

First notice that, in every formula, taking the conjugate of any involved function at point $\mathbf{t}$ is the same as taking the function at point $-\mathbf{t}$. This is also verified for derivatives. Write now for any $\theta \in \Theta_{k^\star}$ and any $\mathbf{t} = (t_1, t_2)$

$$G_n\left(\theta, \mathbf{t}\right) = \widehat{\Phi}_n\left(\mathbf{t}\right) \phi_{\theta,1}\left(t_1\right) \phi_{\theta,2}\left(t_2\right) - \Phi_\theta\left(\mathbf{t}\right) \widehat{\phi}_{n,1}\left(t_1\right) \widehat{\phi}_{n,2}\left(t_2\right)$$

so that, if $\nabla G_n(\theta, \mathbf{t})$ denotes the gradient of G_n with respect to θ at point $(\theta, \mathbf{t})$, one has

$$\nabla M_n(\theta^\star) = \int \left[\nabla G_n(\theta^\star, \mathbf{t}) G_n(\theta^\star, -\mathbf{t}) + \nabla G_n(\theta^\star, -\mathbf{t}) G_n(\theta^\star, \mathbf{t})\right] w(\mathbf{t})\, d\mathbf{t}.$$

Now, writing $\widehat{\Phi}_n(\mathbf{t}) = \frac{Z_n(\mathbf{t})}{\sqrt{n}} + \Phi_{\theta^\star}(\mathbf{t})\phi_{F^\star}(t_1)\phi_{F^\star}(t_2)$ and using (**A2**) one gets easily

$$\begin{aligned}\sqrt{n}\nabla M_n(\theta^\star) = \int & \{\phi_{F^\star}(t_1)\phi_{F^\star}(t_2)\left[\Phi_{\theta^\star}(\mathbf{t})\nabla\left(\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2)\right) - \nabla\Phi_{\theta^\star}(\mathbf{t})\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2)\right] \\ & \left[Z_n(-\mathbf{t})\phi_{\theta^\star}(-t_1)\phi_{\theta^\star}(-t_2) - \Phi_{\theta^\star}(-\mathbf{t})\left(Z_n(-t_1,0)\phi_{\theta^\star}(-t_2) + Z_n(0,-t_2)\phi_{\theta^\star}(-t_1)\right)\right] \\ & +\phi_{F^\star}(-t_1)\phi_{F^\star}(-t_2)\left[\Phi_{\theta^\star}(-\mathbf{t})\nabla\left(\phi_{\theta^\star}(-t_1)\phi_{\theta^\star}(-t_2)\right) - \nabla\Phi_{\theta^\star}(-\mathbf{t})\phi_{\theta^\star}(-t_1)\phi_{\theta^\star}(-t_2)\right] \\ & \left[Z_n(\mathbf{t})\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2) - \Phi_{\theta^\star}(\mathbf{t})\left(Z_n(t_1,0)\phi_{\theta^\star}(t_2) + Z_n(0,t_2)\phi_{\theta^\star}(t_1)\right)\right]\} w(\mathbf{t})\, d\mathbf{t} \\ & + O_{\mathbb{P}^\star}\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

and the convergence in distribution of $\sqrt{n}\nabla M_n(\theta^\star)$ to a centered gaussian with variance V follows.

Similar computation gives that for any $\theta \in \Theta_{k^\star}$

$$\begin{aligned}D_2 M_n(\theta) - D_2 M_n(\theta^\star) &= \int |\widehat{\Phi}_n(\mathbf{t})|^2 \left[A_1(\mathbf{t},\theta) - A_1(\mathbf{t},\theta^\star)\right] w(\mathbf{t}) d\mathbf{t} \\ &+ \int |\widehat{\Phi}_n(t_1,0)|^2 |\widehat{\Phi}_n(0,t_2)|^2 \left[A_2(\mathbf{t},\theta) - A_2(\mathbf{t},\theta^\star)\right] w(\mathbf{t}) d\mathbf{t} \\ &+ Re\left\{\int \widehat{\Phi}_n(-\mathbf{t})\widehat{\Phi}_n(t_1,0)\widehat{\Phi}_n(0,t_2)\left[A_3(\mathbf{t},\theta) - A_3(\mathbf{t},\theta^\star)\right] w(\mathbf{t}) d\mathbf{t}\right\}\end{aligned}$$

for matrix-valued functions $A_1(\mathbf{t},\theta)$, $A_2(\mathbf{t},\theta)$, $A_3(\mathbf{t},\theta)$ that are, in a neighborhood of $\theta^\star$, continuous in the variable θ for all $\mathbf{t}$ and uniformly upper bounded. Thus $D_2M_n(\theta_n) - D_2M_n(\theta^\star)$ converges in $\mathbb{P}^\star$-probability to 0 whenever θ_n is a random variable converging in $\mathbb{P}^\star$-probability to $\theta^\star$.

Finally, note that at point $\theta^\star$ the Hessian of M simplifies into:

$$D_2 M(\theta^\star) = 2 \int H(\mathbf{t}) H(-\mathbf{t})^T \left|\phi_{F^\star}(t_1)\phi_{F^\star}(t_2)\right|^2 w(\mathbf{t}) d\mathbf{t},$$

with

$$H(\mathbf{t}) = \Phi_{\theta^\star}(\mathbf{t})(\phi_{\theta^\star}(t_1)\nabla\phi_{\theta^\star}(t_2) + \nabla\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2)) - \nabla\Phi_{\theta^\star}(\mathbf{t})\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2).$$

Denote by $H_{m_j}(\mathbf{t})$, $j = 2, \ldots, k^\star$, $H_{Q_{j_1,j_2}}(\mathbf{t})$, $j_1, j_2 = 1, \ldots, k^\star$, $(j_1, j_2) \neq (k^\star, k^\star)$ the components of the vector $H(\mathbf{t})$. Definite positiveness of the second derivative of M at $\theta^\star$ can thus be established by proving that, if for all $\mathbf{t} \in \mathcal{S}$,

$$\sum_{j=2}^{k} U_{m_j} H_{m_j}(\mathbf{t}) + \sum_{(j_1,j_2)\neq(k,k)} U_{j_1,j_2} H_{Q_{j_1,j_2}}(\mathbf{t}) = 0 \qquad \text{(A.4)}$$

then

$$U_{m_j} = 0,\ j = 2, \cdots, k^\star,\ U_{j_1,j_2} = 0,\ j_1, j_2 = 1, \ldots, k^\star,\ (j_1, j_2) \neq (k^\star, k^\star).$$

By linear independence of the functions e^{ita} and te^{itb} this implies in particular that for all $\mathbf{t} = (t_1, t_2)$,

$$\begin{aligned}&\sum_{j_1,\cdots,j_4=1}^{k^\star} U_{m_{j_1}} \mu^\star(j_1)\mu^\star(j_2) Q^\star_{j_3,j_4} e^{it_1(m^\star_{j_1}+m^\star_{j_3})+it_2(m^\star_{j_2}+m^\star_{j_4})} \\ &\qquad = \sum_{j_1,\cdots,j_4=1}^{k^\star} U_{m_{j_1}} \mu^\star(j_2)\mu^\star(j_3) Q^\star_{j_1,j_4} e^{it_1(m^\star_{j_1}+m^\star_{j_3})+it_2(m^\star_{j_2}+m^\star_{j_4})} \qquad \text{(A.5)}\end{aligned}$$

with $U_{m_1} = 0$. The smallest possible term $m^\star_{j_1} + m^\star_{j_3}$ with $j_1 > 1$ is equal to $m^\star_2 = m^\star_2 + m^\star_1$ setting $j_1 = 2$ and $j_3 = 1$ only. Thus (A.5) implies that

$$U_{m_2}\mu^\star(2) \sum_{j_2,j_4=1}^{k^\star} \mu^\star(j_2)\, Q^\star_{1,j_4} e^{it_2(m^\star_{j_2}+m^\star_{j_4})} = U_{m_2}\mu^\star(1) \sum_{j_2,j_4=1}^{k^\star} \mu^\star(j_2)\, Q^\star_{2,j_4} e^{it_2(m^\star_{j_2}+m^\star_{j_4})}$$

for all t_2, i.e.

$$U_{m_2}\mu^\star(2)\,\phi_{\theta^\star}(t_2) \sum_{j_4=1}^{k^\star} Q^\star_{1,j_4} e^{it_2 m^\star_{j_4}} = U_{m_2}\mu^\star(1)\,\phi_{\theta^\star}(t_2) \sum_{j_4=1}^{k^\star} Q^\star_{2,j_4} e^{it_2 m^\star_{j_4}}.$$

Since $\phi_{\theta^\star}$ has only isolated zeros this is satisfied if and only if

$$U_{m_2}\mu^\star(2) \sum_{j_4=1}^{k^\star} Q^\star_{1,j_4} e^{it_2 m^\star_{j_4}} = U_{m_2}\mu^\star(1) \sum_{j_4=1}^{k^\star} Q^\star_{2,j_4} e^{it_2 m^\star_{j_4}}.$$

Thus (A.5) is satisfied only if either $U_{m_2} = 0$ or $\mu^\star(2)\, Q^\star_{1,j} = \mu^\star(1)\, Q^\star_{2,j}$ for all j. The latter is impossible since $Q^\star$ is non singular, thus $U_{m_2} = 0$ and (A.5) becomes

$$\begin{aligned}\sum_{j_1=3,j_2,\cdots,j_4=1}^{k^\star} & U_{m_{j_1}}\mu^\star(j_1)\,\mu^\star(j_2)\, Q^\star_{j_3,j_4} e^{it_1(m^\star_{j_1}+m^\star_{j_3})+it_2(m^\star_{j_2}+m^\star_{j_4})} \\ &= \sum_{j_1=3,j_2,\cdots,j_4=1}^{k^\star} U_{m_{j_1}}\mu^\star(j_2)\,\mu^\star(j_3)\, Q^\star_{j_1,j_4} e^{it_1(m^\star_{j_1}+m^\star_{j_3})+it_2(m^\star_{j_2}+m^\star_{j_4})}\end{aligned}$$

The smallest possible value for $m^\star_{j_1} + m^\star_{j_3}$ is then $m^\star_3$ which is obtained with the only configuration $j_1 = 3, j_3 = 1$. The same argument as before leads to $U_{m_3} = 0$. Iteration of the argument leads to $U_{m_j} = 0$ for all $j = 1, \cdots, k^\star$. We now study the derivatives associated to Q. We write U the $k^\star \times k^\star$-matrix whose components are U_{j_1,j_2} for $(j_1, j_2) \neq (k^\star, k^\star)$ and $U_{k^\star,k^\star} = -\sum_{(j_1,j_2)\neq(k^\star,k^\star)} U_{j_1,j_2}$. Then

$$\sum_{(j_1,j_2)\neq(k^\star,k^\star)} U_{j_1,j_2} \nabla_{Q_{j_1,j_2}} \Phi_{\theta^\star}(\mathbf{t}) = V(t_1)^T U V(t_2)$$

where for any $t \in \mathbb{R}$, $V(t) = ((e^{itm^\star_j})_{j=1,\cdots,k^\star})^T$, and

$$\sum_{(j_1,j_2)\neq(k^\star,k^\star)} U_{j_1,j_2} \nabla_{Q_{j_1,j_2}} \phi_{\theta^\star}(t_1) = V(t_1)^T U \mathbb{1}$$

with $\mathbb{1} = (1, \cdots, 1)^T \in \mathbb{R}^{k^\star}$, since $\phi_{\theta^\star}(t_1) = V(t_1)^T Q^\star \mathbb{1}$ and $\Phi_{\theta^\star}(\mathbf{t}) = V(t_1)^T Q^\star V(t_2)$. We can then express (A.4) as

$$\begin{aligned}V(t_1)^T \big[& Q^\star V(t_2)V(t_2)^T U \mathbb{1}\mathbb{1}^T (Q^\star)^T + Q^\star V(t_2)V(t_2)^T Q^\star \mathbb{1}\mathbb{1}^T U^T \\ & - U V(t_2)V(t_2)^T Q \mathbb{1}\mathbb{1}^T (Q^\star)^T \big] V(t_1) = 0. \quad (A.6)\end{aligned}$$

Note also that since all differences $m^\star_{j_1} - m^\star_{j_2}$, $j_1 \neq j_2$, are distinct, if A is a $k^\star \times k^\star$-matrix and $\mathcal{I}$ is an open subset of $\mathbb{R}$,

$$\big[\forall t \in \mathcal{I},\ V(t)^T A V(t) = 0\big] \Longrightarrow A + A^T = 0. \quad (A.7)$$

Then (A.6) implies

$$\begin{aligned}& Q^\star V(t_2)V(t_2)^T U \mathbb{1}\mathbb{1}^T (Q^\star)^T + Q^\star \mathbb{1}\mathbb{1}^T U^T V(t_2)V(t_2)^T (Q^\star)^T \\ &\quad + Q^\star V(t_2)V(t_2)^T Q^\star \mathbb{1}\mathbb{1}^T U^T + U \mathbb{1}\mathbb{1}^T (Q^\star)^T V(t_2)V(t_2)^T (Q^\star)^T \\ &\quad - U V(t_2)V(t_2)^T Q^\star \mathbb{1}\mathbb{1}^T (Q^\star)^T - Q^\star \mathbb{1}\mathbb{1}^T (Q^\star)^T V(t_2)V(t_2)^T U^T = 0. \quad (A.8)\end{aligned}$$

Recall also that $\mathbb{1}^T U \mathbb{1} = 0$ and that $Q^\star \mathbb{1} = \mu^\star$. Note that $U\mathbb{1} = \alpha\mu^\star$ with $\alpha \in \mathbb{R}$ if and only if $\alpha = 0$ since $\mathbb{1}^T U\mathbb{1} = 0$ while $\mathbb{1}^T\mu^\star = 1$. Therefore if $U\mathbb{1} \neq 0$ there exists $w \in \mathbb{R}^{k^\star}$ such that $w^T(U\mathbb{1}) \neq 0$ while $(\mu^\star)^T w = 0$. Multiplying the above equality on the left by w^T and on the right by w leads to

$$w^T Q^\star V(t_2)V(t_2)^T(\mu^\star)(U\mathbb{1})^T w = 0$$

that for all t_2 in an open set. Using (A.7) again and since $(U\mathbb{1})^T w \neq 0$ we get that

$$\mu^\star[(Q^\star)^T w]^T + [(Q^\star)^T w](\mu^\star)^T = 0.$$

Since $\mu^\star(j) > 0$ for all j this implies that $(Q^\star)^T w = 0$ which is impossible since $Q^\star$ has full rank. Therefore $U\mathbb{1} = 0$ and (A.8) becomes $V(t_2)^T\mu^\star[UV(t_2)(\mu^\star)^T + \mu^\star V(t_2)^T U^T] = 0$, that is $UV(t_2)(\mu^\star)^T + \mu^\star V(t_2)^T U^T = 0$ for all t_2 in an open set. Multiplying on the left by $\mathbb{1}$ implies that $UV(t_2) = 0$ for all t_2 in an open set so that $U = 0$. □

B Proof of Theorem 3.2

Define for any $\theta \in \Theta_{k^\star}$, $L_n(\theta) = M_n(\theta) - M(\theta)$. Then, since $M_n(\overline{\theta}_n(\mathcal{K})) \leq M_n(\theta^\star)$, one easily gets

$$M\left(\overline{\theta}_n(\mathcal{K})\right) - M\left(\theta^\star\right) \leq \left|L_n\left(\overline{\theta}_n(\mathcal{K})\right) - L_n\left(\theta^\star\right)\right|.$$

Define for any $\mathbf{t} = (t_1, t_2)$ and any θ

$$G\left(\theta, \mathbf{t}\right) = \left\{\Phi_{\theta^\star}\left(\mathbf{t}\right)\phi_{\theta,1}\left(t_1\right)\phi_{\theta,2}\left(t_2\right) - \Phi_\theta\left(\mathbf{t}\right)\phi_{\theta^\star,1}\left(t_1\right)\phi_{\theta^\star,2}\left(t_2\right)\right\}\phi_{F^\star}\left(t_1\right)\phi_{F^\star}\left(t_2\right)$$

and

$$\begin{aligned} B_n\left(\theta, \mathbf{t}\right) = \phi_{F^\star}\left(t_1\right)\phi_{F^\star}\left(t_2\right)\Bigg\{&\frac{Z_n(\mathbf{t})}{\sqrt{n}}\phi_{\theta,1}\left(t_1\right)\phi_{\theta,2}\left(t_2\right) \\ &-\Phi_\theta\left(\mathbf{t}\right)\left[\frac{Z_n(t_1,0)}{\sqrt{n}}\phi_{\theta,2}\left(t_2\right) + \frac{Z_n(0,t_2)}{\sqrt{n}}\phi_{\theta,1}\left(t_1\right) + \frac{Z_n(t_1,0)Z_n(0,t_2)}{n}\right]\Bigg\}\end{aligned}$$

Writing $\widehat{\Phi}_n\left(\mathbf{t}\right) = \frac{Z_n(\mathbf{t})}{\sqrt{n}} + \Phi_{\theta^\star}(\mathbf{t})\phi_{F^\star}(t_1)\phi_{F^\star}(t_2)$ one gets

$$L_n\left(\theta\right) = \int \left(\left[B_n\left(\theta, \mathbf{t}\right) + G\left(\theta, \mathbf{t}\right)\right]\left[B_n\left(\theta, -\mathbf{t}\right) + G\left(\theta, -\mathbf{t}\right)\right] - \left|G\left(\theta, \mathbf{t}\right)\right|^2\right) w\left(\mathbf{t}\right) d\mathbf{t}.$$

Since $G\left(\theta^\star, \mathbf{t}\right) = 0$ for all $\mathbf{t}$ we obtain

$$\begin{aligned} L_n\left(\theta\right) - L_n\left(\theta^\star\right) = \int \Big\{&\left|B_n\left(\theta, \mathbf{t}\right)\right|^2 - \left|B_n\left(\theta^\star, \mathbf{t}\right)\right|^2 + B_n\left(\theta, \mathbf{t}\right)G\left(\theta, -\mathbf{t}\right) \\ &+ B_n\left(\theta, -\mathbf{t}\right)G\left(\theta, \mathbf{t}\right)\Big\} w\left(\mathbf{t}\right) d\mathbf{t}\end{aligned}$$

which gives

$$\begin{aligned} \left|L_n\left(\theta\right) - L_n\left(\theta^\star\right)\right| \leq \int \Big\{&\left|B_n\left(\theta, \mathbf{t}\right) - B_n\left(\theta^\star, \mathbf{t}\right)\right|\left|B_n\left(\theta, \mathbf{t}\right) + B_n\left(\theta^\star, \mathbf{t}\right)\right| \\ &+ 2\left|B_n\left(\theta, \mathbf{t}\right)\right|\left|G\left(\theta, \mathbf{t}\right) - G\left(\theta^\star, \mathbf{t}\right)\right|\Big\} w\left(\mathbf{t}\right) d\mathbf{t}\end{aligned}$$

which leads to

$$M\left(\overline{\theta}_n(\mathcal{K})\right) - M\left(\theta^\star\right) \leq CW_n\|\overline{\theta}_n(\mathcal{K}) - \theta^\star\| \tag{B.1}$$

for some constant C and any integer n, and with

$$W_n = \left\{\frac{V_n}{\sqrt{n}} + \frac{V_n^2}{n} + \frac{V_n^3}{n^{3/2}} + \frac{V_n^4}{n^2}\right\},\ V_n = \sup_{\mathbf{t}\in\mathcal{S}}\left|Z_n\left(\mathbf{t}\right)\right|.$$

Observe now that, since D_2M is continuous and $D_2M(\theta^\star)$ is non singular, there exists $\lambda > 0$ and $\alpha > 0$ such that, if $\|\theta - \theta^\star\| \leq \alpha$, then $M(\theta) - M(\theta^\star) \geq \frac{\lambda}{2}\|\theta - \theta^\star\|^2$. Moreover, there exists $\delta > 0$ such that, if $\theta \in \mathcal{K}$ is such that $\|\theta - \theta^\star\| \geq \alpha$, then $M(\theta) - M(\theta^\star) \geq \delta$. Using (B.1) we obtain that for any real number M large enough,

$$\mathbb{P}^\star\left(\sqrt{n}\|\overline{\theta}_n(\mathcal{K}) - \theta^\star\| \geq M\right) \leq \mathbb{P}^\star\left(W_n \geq \frac{\delta}{2CM(\mathcal{K})}\right) + \mathbb{P}^\star\left(\sqrt{n}W_n \geq \frac{M\lambda}{2C}\right)$$

where $M(\mathcal{K}) = \sup_{\theta\in\mathcal{K}} \|\theta\|$. This last equation together with Assumption (**A3**) gives the Theorem.

C Proof of Theorem 4.1

The proof follows the general methodology for model selection developed by Massart (2007). To prove Theorem 4.1 and Theorem 4.2, we will use a concentration inequality we state now. Let us introduce some notations. For any real function f, denote

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[f(Y_i) - \int f d\mathbb{P}^\star\right].$$

Lemma 3 *Assume* (**A4**). *Let $\mathcal{F}$ be a class of real functions, and F such that, for any $f \in \mathcal{F}$, $|f| \leq F$. Assume that there exists $c(F) > 0$ and $C(F) > 0$ such that $\forall j = 1, \dots, k^\star$, $|g(j)| \leq C(F)$ where g is defined by*

$$g(j) = \ln E_{\mathbb{P}^\star}\left\{\exp\left[2c(F)^{-1}|F(Y_2)|\right] |S_1 = j\right\}.$$

Then there exist universal constants C_1, C_2, K_1, K_2 and a constant $C^\star$ depending only on $Q^\star$ such that

$$\mathbb{P}^\star\left(\sqrt{n}\sup_{f\in\mathcal{F}} \mathbb{G}_n f \geq K_1\sqrt{n}E_{\mathbb{P}^\star}\left(\sup_{f\in\mathcal{F}} \mathbb{G}_n f\right) + C_1\tau\sqrt{nx} + C_2C^\star c(F)C(F)x\right) \leq K_2 \exp\{-x\}$$

where $\tau^2 = \sup_{f\in\mathcal{F}} E_{\mathbb{P}^\star} f^2(Y_1)$.

Proof of Lemma 3

The lemma is an application of Theorem 7 in Adamczak and Bednorz (2012) to the stationary Markov chain $(X_i)_{i\geq1} = (S_i, Y_i)_{i\geq1}$ and functions $f(s,y) := f(y)$. Then, with the notations of Adamczak and Bednorz (2012) we get that:

- $m = 1$,
- the small set C is the whole space,
- the minorizing probability measure ν is that of $(\tilde{S}_i, Y_i)_{i\geq1}$ with $(\tilde{S}_i)_i$ i.i.d. with uniform distribution, and $\delta = \min_{i,j} Q^\star_{i,j}$.
- Since C is the whole space, the return times $\sigma(i) = i$, so that $s_i(f) = f(Y_i)$, thus the σ^2 of Theorem 7 is just $\sup_f E_{\mathbb{P}^\star}(f^2(Y_1))$.

Using the specific assumption of the lemma, taking $\alpha = 1$, we can apply Corollary 1 of Adamczak and Bednorz (2012), to get (with their notations again)

$$a, b, c \leq C^\star c(F)C(F)$$

for some constant $C^\star > 0$ depending only on $\min_{i,j} Q^\star_{i,j}$. □

For any $p \geq 2$, define

$$S_p = \left\{ \sum_{j=1}^{k^\star} \mu(j) f(\cdot - m_j), f \in \mathcal{F}_p, \ |m_j| \leq M(\mathcal{K}), \ \sum_{j=1}^{k^\star} \mu(j) = 1, \right.$$
$$\left. \mu(j) \geq 0, \ j = 1, \ldots, k^\star \right\},$$

so that $\widehat{s}_p \in S_p$. We now fix, for any $p \geq 2$, some $\tilde{s}_p \in S_p$ such that:

$$\forall t \in S_p, \ \|\sqrt{s^\star} - \sqrt{\tilde{s}_p}\|^2 \leq 2\|\sqrt{s^\star} - \sqrt{t}\|^2. \tag{C.1}$$

For any $p \geq 2$ and any $\sigma > 0$, define

$$W_p(\sigma) = \sup_{t \in \mathcal{S}_p, \|\sqrt{t} - \sqrt{\tilde{s}_p}\|_2 \leq \sigma} \mathbb{G}_n \left(\ln \left(\frac{s^\star + t}{s^\star + \tilde{s}_p} \right) \right),$$

and let L_p be an enveloppe function of $\{\ln(s^\star + t) - \ln(s^\star + \tilde{s}_p), t \in \mathcal{S}_p\}$. Assume there exists functions ψ_p such that $\psi_p(x)/x$ is non increasing and for all $p \geq 2$ and $\sigma > 0$,

$$E_{\mathbb{P}^\star}[W_p(\sigma)] \leq \psi_p(\sigma). \tag{C.2}$$

Define σ_p (depending also on n) as the unique solution of

$$\psi_p(\sigma_p) = \sqrt{n}\sigma_p^2. \tag{C.3}$$

Now we follow and adapt the proof of Theorem 7.11 in Massart (2007). Let p be such that $K(s^\star, s_p) < +\infty$. If p' is such that $D(p') \leq D(p)$, then one gets, as in Massart (2007) p.241,

$$\begin{aligned} K\left(s^\star, \frac{s^\star + \widehat{s}_{p'}}{2}\right) &- \frac{1}{2\sqrt{n}} \mathbb{G}_n \left(\log \left(\frac{s_p}{s^\star} \right) \right) \\ &\leq K(s^\star, s_p) + \mathrm{pen}(p, n) - \frac{1}{\sqrt{n}} \mathbb{G}_n \left(\ln \left(\frac{s^\star + \widehat{s}_{p'}}{2s^\star} \right) \right) - \mathrm{pen}(p', n) + V_p \end{aligned} \tag{C.4}$$

where

$$V_p = \frac{1}{n} \sum_{i=1}^n \ln \left(\frac{\sum_{j=1}^{k^\star} \widehat{\mu}(j) f_p(Y_i - \widehat{m}_j)}{\sum_{j=1}^{k^\star} \mu^\star(j) f_p(Y_i - m_j^\star)} \right).$$

Applying Lemma 4.23 in Massart (2007) p. 139, for any positive $y_{p'}$:

$$E^\star \left[\sup_{t \in \mathcal{S}_{p'}} \mathbb{G}_n \left(\frac{\ln(s^\star + t) - \ln(s^\star + \tilde{s}_{p'})}{y_{p'}^2 + \|\sqrt{t} - \sqrt{s_{p'}}\|_2^2} \right) \right] \leq 4 \frac{\psi_{p'}(y_{p'})}{y_{p'}^2}.$$

Using Lemma 3, the fact that $2y_{p'}\|\sqrt{t} - \sqrt{\tilde{s}_{p'}}\|_2 \leq y_{p'}^2 + \|\sqrt{t} - \sqrt{\tilde{s}_{p'}}\|_2^2$, and Lemma 7.26 p. 276 in Massart (2007), we obtain that for some constant $C > 0$, except on a set with probability less than $K_2 \exp -(x_{p'} + x)$, for all $x > 0$:

$$\frac{1}{\sqrt{n}} \mathbb{G}_n \left(\frac{\ln(s^\star + \widehat{s}_{p'}) - \ln(s^\star + \tilde{s}_{p'})}{y_{p'}^2 + \|\sqrt{\tilde{s}_{p'}} - \sqrt{\widehat{s}_{p'}}\|^2} \right) \leq \frac{Cte}{y_{p'}} \left(\frac{\psi_{p'}(y_{p'})}{y_{p'}\sqrt{n}} + \frac{T(L_{p'})(x_{p'} + x)}{ny_{p'}} + \sqrt{\frac{x_{p'} + x}{n}} \right).$$

Here, $T(L_{p'}) = C_2 C^\star c(L_{p'}) C(L_{p'})$. Using again Lemma 3 and Lemma 7.26 p. 276 in Massart (2007) we get that, for some constant $C > 0$, except on a set with probability less than $K_2 \exp -(x_{p'} + x)$, for all $x > 0$:

$$\frac{1}{\sqrt{n}} \frac{\mathbb{G}_n(\ln(s^\star + \tilde{s}_{p'}) - \ln(2s^\star))}{y_{p'}^2 + \|\sqrt{s^\star} - \sqrt{\tilde{s}_{p'}}\|^2} \leq \frac{C}{y_{p'}} \left(\frac{T(L_{p'})(x_{p'} + x)}{ny_{p'}} + \sqrt{\frac{x_{p'} + x}{n}} \right).$$

Now, using (C.1), we get

$$\|\sqrt{\tilde{s}_{p'}}-\sqrt{\widehat{s}_{p'}}\|^2 \leq \left[\|\sqrt{\tilde{s}_{p'}}-\sqrt{s^\star}\|+\|\sqrt{s^\star}-\sqrt{\widehat{s}_{p'}}\|\right]^2 \leq 6\|\sqrt{s^\star}-\sqrt{\widehat{s}_{p'}}\|^2$$

and

$$\|\sqrt{s^\star}-\sqrt{\tilde{s}_{p'}}\|^2 \leq 2\|\sqrt{s^\star}-\sqrt{\widehat{s}_{p'}}\|^2$$

and we finally obtain that, for some other constant $C>0$ depending only on $\mathbb{P}^\star$, except on a set with probability less than $2K_2\exp-(x_{p'}+x)$, for all $x>0$:

$$\frac{1}{\sqrt{n}}\mathbb{G}_n\left(\frac{\ln(s^\star+\widehat{s}_{p'})-\ln(2s^\star)}{y_{p'}^2+\|\sqrt{s^\star}-\sqrt{\widehat{s}_{p'}}\|^2}\right) \leq \frac{C}{y_{p'}}\left(\frac{\psi_{p'}(y_{p'})}{y_{p'}\sqrt{n}}+\frac{T(L_{p'})(x_{p'}+x)}{ny_{p'}}+\sqrt{\frac{x_{p'}+x}{n}}\right).$$

Define for some constant a to be chosen

$$y_{p'} = a^{-1}\sqrt{\sigma_{p'}^2+\frac{(x_{p'}+x)(1+T(L_{p'}))}{n}}.$$

Then we can follow the proof of Theorem 7.11 in Massart (2007) to obtain that, as soon as

$$\operatorname{pen}(p,n) \geq \kappa\left(\sigma_p^2+\frac{x_p(1+T(L_p))}{n}\right), \tag{C.5}$$

one has for any $n\geq 2$, for some real numbers $\kappa>0$ and $C>0$ depending only on $Q^\star$

$$E_{\mathbb{P}^\star}\left[h^2(s^\star,\widehat{s}_{\widehat{p}})\right] \leq C\left\{\inf_{p\geq 2}\left(K(s^\star,s_p)+\operatorname{pen}(p,n)+E_{\mathbb{P}^\star}[V_p]\right)+\frac{\Sigma}{n}\right\}.$$

But using the convexity of the Kullback-Leibler divergence to both arguments, we have, for any $p\geq 2$, $K(s^\star,s_p)\leq K(f^\star,f_p)$. Thus to finish the proof of Theorem 4.1, one has to find functions ψ_p verifying (C.2), evaluate σ_p using (C.3), and evaluate $T(L_p)$.
Let us first prove that there exists constants $C, C'>0$ depending only on δ and $Q^\star$ such that, as soon as (**A4**) holds, for any $p\geq 2$,

$$T(L_p) \leq C\ln\left(1+\frac{C'}{b_p^\delta}\right). \tag{C.6}$$

First of all, we see that we can take

$$L_p(y) = \ln\left(1+\frac{1}{b_p\sqrt{2\pi}s^\star(y)}\right),$$

with $c(L_p)=2/\delta$, the function defined in Lemma 3 is given by

$$g(s) = \log\left[\sum_{j=1}^{k^\star}Q_{s,j}^\star\int\left(1+\frac{1}{b_p\sqrt{2\pi}s^\star(u)}\right)^\delta f^\star(u-m_j^\star)du\right]$$

Under (**A4**), on gets that there exists constants $C>0$ depending only in $Q^\star$ and δ such that g is bounded by the constant $C\ln\left(1+\frac{C'}{b_p^\delta}\right)$ and (C.6) follows (for maybe another constant C).
To find functions ψ_p, we shall use Doukhan et al. (1995). Since $(Y_t)_{t\in\mathbb{N}}$ is geometrically ergodic, Lemma 2 in Doukhan et al. (1995), implies that, for some constant C that depends only on $Q^\star$, for any real function f,

$$\|f\|_\beta^2 \leq C\gamma(f)(1+\log^+(\gamma(f))), \quad \gamma(f)=\int f^2(1+\log^+|f|)d\mathbb{P}^\star$$

where $\|\cdot\|_\beta$ is defined in Doukhan et al. (1995). Now, since for all $x>0$, $x\ln^+ x \le x^2/e$,

$$\|f\|_\beta^2 \le \frac{C}{e}\int |f|^3 d\mathbb{P}^\star \left(1+\log^+(\frac{1}{e}\int |f|^3 d\mathbb{P}^\star)\right).$$

Using Lemma 7.26 in Massart (2007), we thus get that for all $t\in\mathcal{S}_p$,

$$\|\ln(s^\star+t)-\ln(s^\star+\tilde{s}_p)\|_\beta^2 \quad\le\quad \frac{1}{c^2}\|\sqrt{t}-\sqrt{\tilde{s}_p}\|_2^2$$

for some constant $c>0$ that depends only on $Q^\star$, and the same trick leads to

$$H_\beta\left(u,\{\ln(s^\star+t)-\ln(s^\star+\tilde{s}_p)\,,\ t\in\mathcal{S}_p\}\right)\le H_2\left(cu,\{\sqrt{t},\ t\in\mathcal{S}_p\}\right)$$

where $H_\beta(u,\mathcal{F})$ is the bracketing entropy of a set $\mathcal{F}$ at level u with respect to $\|\cdot\|_\beta$, that is the logarithm of the minimum of the number of brackets of $\|\cdot\|_\beta$-width u needed to cover $\mathcal{F}$, and $H_2(u,\mathcal{F})$ is the bracketing entropy of a set $\mathcal{F}$ at level u with respect to $\|\cdot\|_2$.
Let for any for $\sigma>0$ and $p\ge 2$

$$\eta_p(\sigma)=\int_0^{\sigma/c}\sqrt{H_2\left(cu,\{\sqrt{t},\ t\in\mathcal{S}_p\}\right)}du.$$

Using Theorem 3 in Doukhan et al. (1995) we get

$$E_{\mathbb{P}^\star}[W_p(\sigma)]\le A\eta_p(\sigma)\left[1+\frac{\delta_p(1\wedge\epsilon(\sigma,n))}{\sigma}\right], \tag{C.7}$$

where $\epsilon(\sigma,n)$ is the unique solution of $x^2/B(x)=\eta_p^2(\sigma)/n\sigma^2$,

$$B(x)=x+C(x-x\ln x)$$

for some constant C that depends only on $Q^\star$, and δ_p is the function given by

$$\delta_p(\epsilon)=\sup_{t\le\epsilon}Q(t)\sqrt{B(t)}$$

with for any t, $Q(t)\le u$ iff $\mathbb{P}^\star(H_p(Y_1)>u)\le t$. Here, H_p is an envelope function of $\{\ln(s^\star+t)-\ln(s^\star+\tilde{s}_p)\,,\ \|\sqrt{t}-\sqrt{\tilde{s}_p}\|\le\sigma,\ t\in\mathcal{S}_p\}$. Taking $H_p=L_p$ one gets easily

$$Q(t)\le\ln\left(1+\frac{1}{tb_p\sqrt{2\pi}}\right),$$

so that $\delta_p(\epsilon)\le\sup_{t\le\epsilon}h_p(t)$ with

$$h_p(t)=\ln\left(1+\frac{1}{tb_p\sqrt{2\pi}}\right)\sqrt{t+C(t-t\ln t)}.$$

The variations of h_p imply that there exists a universal constant b such that as soon as $b_p\le b$, $h(t)$ is increasing on $(0,1)$, so that

$$\delta_p(\epsilon\wedge 1)=h_p(\epsilon\wedge 1)\le\tilde{h}_p(\epsilon\wedge 1)$$

with

$$\tilde{h}_p(t)=C\ln\left(\frac{1}{b_p}\right)\sqrt{t}|\ln t|\left(\sqrt{|\ln t|}\wedge 1\right),$$

for some universal constant C. Using Maugis and Michel (2011), we get that for some fixed constant K, for all $u>0$,

$$H_2\left(u,\{\sqrt{t},\ t\in\mathcal{S}_p\}\right)\le k^\star p\left[3\ln\left(\frac{1}{u\wedge 1}\right)+\frac{3}{4}\ln\left(\frac{1}{b_p}\right)+\ln A_p+K\right]+\ln(k^\star p).$$

Using the fact that for all $\epsilon \in]0,1]$,

$$\int_0^\epsilon \sqrt{\ln\left(\frac{1}{x}\right)}dx \leq \epsilon \left\{\sqrt{\ln\left(\frac{1}{\epsilon}\right)} + \sqrt{\pi}\right\}$$

and since

$$\eta_p(\sigma) = \frac{1}{c}\int_0^\sigma \sqrt{H_2\left(c^2u, \{\sqrt{t},\ t \in \mathcal{S}_p\}\right)}du$$

we get that for some other fixed constant K and all $\sigma > 0$

$$\eta_p(\sigma) \leq \frac{\sigma}{c} t_p(\sigma) \tag{C.8}$$

with

$$t_p(\sigma) = \sqrt{k^\star p}\left[3\sqrt{\ln\left(\frac{1}{\sigma \wedge 1}\right)} + \sqrt{\frac{3}{4}\ln\left(\frac{1}{b_p}\right) + \ln A_p + K}\right] + \sqrt{\ln(k^\star p)}.$$

Now, one may use the upper bound (C.8) to upper bound $\epsilon(\sigma, n)$, and we get that for some universal constant C,

$$\epsilon(\sigma, n) \leq C\frac{t_p^2(\sigma)}{n}\ln\left[\frac{t_p^2(\sigma)}{2n}\right].$$

Then we may set

$$\psi_p(\sigma) = \frac{\sigma}{c}t_p(\sigma)\left[1 + \frac{\tilde{h}\left(C\frac{t_p^2(\sigma)}{n}\ln\left[\frac{t_p^2(\sigma)}{2n}\right] \wedge 1\right)}{\sigma}\right].$$

(C.2) holds, $\psi_p(x)/x$ is indeed non increasing, and if σ_p is the unique solution of (C.3), we obtain that for some constant C depending only on $\mathbb{P}^\star$, as soon as $b_p \leq b$,

$$\sigma_p^2 \geq \frac{C}{n}k^\star p\left[\ln n + \ln\left(\frac{1}{b_p}\right) + \ln A_p\right]. \tag{C.9}$$

D Proof of Theorem 4.2

For simplicity's sake we denote in the following $\mathcal{H}_{loc}(\beta) := \mathcal{H}_{loc}(\beta, \gamma, \mathcal{P})$. Set $p = p_0\lfloor (n/\log n)^{1/(2\beta+1)}(\log n)^{4\beta/(2\beta+1}$ with $p_0 > 0$ fixed which we shall determine later, $b_p = b_0(\log p)^2/p$ for some positive b_0 and $A_p = a_0|\log b_p|$ for some positive a_0. The approximating $f_p \in \mathcal{F}_p$

$$f_p(y) = \sum_{i=1}^p \pi_i\varphi_{b_p}(y - \alpha_i)$$

is taken from Kruijer et al. (2010). Let $\ell_j^\star$ denote the j-th derivative of $\log f^\star$. A simple modification in the proof of Lemma 4 of Kruijer et al. (2010) gives that for any H and any $\tilde{H}$ with $H > \tilde{H} + 3\beta$, there exists $\tilde{B}$ such that if

$$\begin{aligned} D_p := \Big\{y\ :\ & f^\star(y - m) \geq b_p^{\tilde{H}}, |\ell_j^\star(y - m)| \leq \tilde{B}b_p^{-j}|\log p|^{-j/2},\ j \leq \beta, \\ & |L(y - m)| \leq \tilde{B}b_p^{-\beta}|\log p|^{-\beta/2},\ \forall 0 \leq m \leq 2m_k^\star\Big\} \end{aligned}$$

then, for all $y \in D_p$ and all $0 \leq m \leq m_k^\star$

$$f_p(y - m) = f^\star(y - m)(1 + O(R(y - m)b_p^\beta)) + O((1 + R(y - m))b_p^{H-\tilde{H}}), \tag{D.1}$$

where the function $R(y)$ is a linear combination of $L(y)$ and of the functions $|\ell_j^\star(y)|^{\beta/j}$, $j \le \beta$, and where the constants entering the terms $O(.)$ depend on $\mathcal{H}_{loc}(\beta)$, $\tilde{B}$, H and $\tilde{H}$. Note that since the functions $l_j^\star$ are bounded by polynomials, there exists a constant C such that $|R(y-m)| \le C(1+R(y))$, $\forall 0 \le m \le 2m_k^*$. In the following we fix $\tilde{H} > 4\beta + 2\gamma$ and $H > \tilde{H} + 3\beta$. Moreover, Lemma 4 in Kruijer et al. (2010) implies

$$K(f^\star, f_p) \lesssim b_p^{2\beta}, \qquad \int f^\star \left(\log f^\star - \log f_p\right)^2(y)dy \lesssim b_p^{2\beta}. \tag{D.2}$$

Here and further, $\lesssim$ will denote an upper bound up to a constant, where the constant entering the upper bound depends only on $\mathcal{H}_{loc}(\beta)$. Throughout the proof C denotes a generic constant depending only on $H_{loc}(\beta)$ and $Q^\star$.

First of all, with such choices of p, b_p, and A_p, using Theorem 4.1 and (D.2), there remains to prove that $E_{\mathbb{P}^\star}[V_p] \lesssim b_p^{2\beta}$ or equivalently $v_n E_{\mathbb{P}^\star}[V_p] \lesssim 1$ with $v_n = n^{\frac{2\beta}{2\beta+1}}(\log n)^{-6\beta/(2\beta+1)}$. For any θ and any y, set

$$w_p(\theta, y) = \log\left(\frac{\sum_{j=1}^{k^\star} \mu(j) f_p(y - m_j)}{\sum_{j=1}^{k^\star} \mu^\star(j) f_p(y - m_j^\star)}\right).$$

First note that

$$\log\left(\frac{\sum_{j=1}^{k} \mu_j^\star f_p(y - m_j)}{\sum_{j=1}^{k} \mu_j^\star f_p(y - m_j^\star)}\right) \le \max_j \left\{\frac{(|y - m_j^\star| + A_p)|m_j - m_j^\star|}{b_p^2}\right\}. \tag{D.3}$$

Thus we can bound

$$\begin{aligned}
&\frac{v_n}{n} E_{\mathbb{P}^\star}\left[\sum_{i=1}^n w_p(\widehat{\theta}, Y_i)\mathbb{1}_{\|\theta - \widehat{\theta}\| > M_0\sqrt{\log n/n}}\right] \\
&\le \frac{v_n}{n} \max_j \sum_{i=1}^n E_{\mathbb{P}^\star}\left[\mathbb{1}_{\|\theta - \widehat{\theta}\| > M_0\sqrt{\log n/n}}\left(\frac{(|Y_i - m_j^\star| + A_p)|\hat{m}_j - m_j^\star|}{b_p^2} + \frac{|\hat{\mu}_j - \mu_j^\star|}{\mu_j^\star}\right)\right] \\
&\lesssim \left(\frac{v_n \log p}{b_p^2\sqrt{n}} + \frac{v_n}{\sqrt{n}}\right)\mathbb{P}^\star\left[\|\theta - \widehat{\theta}\| > M_0\sqrt{\log n/n}\right] \\
&= o(1)
\end{aligned}$$

by Theorem 3.2 and choosing $M_0 = 1/\sqrt{c^\star}$.
Set now $H_1 > 3 + 2\beta$, $C_{p,1} = D_p^c \cap \{|y| \le H_1 \log(1/b_p)\tau^{-1}\}$ and $C_{p,2} = D_p^c \cap \{|y| > H_1 \log(1/b_p)\tau^{-1}\}$. Using (D.3) we get, for all $i = 1, \cdots, n$,

$$\begin{aligned}
E_{\mathbb{P}^\star}\left[\mathbb{1}_{C_{p,1}}(Y_i) w_p(\widehat{\theta}, Y_i)\mathbb{1}_{\|\theta - \widehat{\theta}\| \le M_0\sqrt{\log n/n}}\right] &\lesssim \frac{(\log p)^{3/2}}{b_p^2\sqrt{n}} \int_{C_{p,1}} s^\star(y)dy \\
&\le \frac{(\log p)^{3/2}}{b_p^2\sqrt{n}} b_p^{2\beta+\gamma} \lesssim v_n^{-1}
\end{aligned}$$

as soon as $\gamma > (3/2 - \beta)_+$, where the last inequality comes from an adaptation of Lemma 2 in Kruijer et al. (2010), using the moment conditions (4.2). We also have

$$\begin{aligned}
E_{\mathbb{P}^\star}\left[\mathbb{1}_{C_{p,2}}(Y_i) w_p(\widehat{\theta}, Y_i)\mathbb{1}_{\|\theta - \widehat{\theta}\| \le M_0\sqrt{\log n/n}}\right] &\lesssim \frac{(\log p)^{3/2}}{b_p^2\sqrt{n}} \int_{C_{p,2}} |y| s^\star(y)dy \\
&\lesssim \frac{(\log p)^{3/2} b_p^{H_1/2}}{b_p^2\sqrt{n}} \lesssim v_n^{-1}
\end{aligned} \tag{D.4}$$

since $H_1 > 3 + 2\beta$, where the last inequality comes from the tail condition (4.2). There thus remains to prove that

$$v_n E_{\mathbb{P}^\star}\left[\frac{1}{n}\sum_{i=1}^n w_p(\widehat{\theta}_n, Y_i)\mathbb{1}_{D_p}(Y_i)\mathbb{1}_{\|\theta - \widehat{\theta}\| \le M_0\sqrt{\log n/n}}\right] \lesssim 1.$$

We shall use:

$$v_n E_{\mathbb{P}^\star}\left[\frac{1}{n}\sum_{i=1}^n w_p(\widehat{\theta}_n, Y_i)1\!\!1_{D_p}(Y_i)1\!\!1_{\|\theta-\widehat{\theta}\|\le M_0\sqrt{\log n/n}}\right]$$
$$\le \int_0^{+\infty}\mathbb{P}^\star\left(\frac{v_n}{n}\sum_{i=1}^n w_p(\widehat{\theta}_n, Y_i)1\!\!1_{D_p}(Y_i)1\!\!1_{\|\theta-\widehat{\theta}\|\le M_0\sqrt{\log n/n}} \ge x\right)dx. \quad \text{(D.5)}$$

Notice now that

$$\mathbb{P}^\star\left(\frac{v_n}{n}\sum_{i=1}^n w_p(\widehat{\theta}_n, Y_i)1\!\!1_{D_p}(Y_i)1\!\!1_{\|\theta-\widehat{\theta}\|\le M_0\sqrt{\log n/n}} \ge x\right) \le$$
$$\mathbb{P}^\star\left(\sqrt{n}\|\widehat{\theta}_n-\theta^\star\|\ge M(x)\right)+\mathbb{P}^\star\left(\sup_{\sqrt{n}\|\theta-\theta^\star\|\le M(x)\wedge M_0\sqrt{\log n/n}}\frac{v_n}{\sqrt{n}}\mathbb{G}_n(w_p(\theta,\cdot)1\!\!1_{D_p}(\cdot))\ge\frac{x}{2}\right)$$

as soon as

$$v_n \sup_{\sqrt{n}\|\theta-\theta^\star\|\le M(x)\wedge M_0\sqrt{\log n}} E_{\mathbb{P}^\star}[w_p(\theta,Y_1)1\!\!1_{D_p}(Y_1)]\le\frac{x}{2}. \quad \text{(D.6)}$$

If moreover

$$\frac{v_n K_1}{\sqrt{n}}E_{\mathbb{P}^\star}\left(\sup_{\sqrt{n}\|\theta-\theta^\star\|\le M(x)\wedge M_0\sqrt{\log n}}\mathbb{G}_n(w_p(\theta,\cdot)1\!\!1_{D_p}(\cdot))\right)\le\frac{x}{4}, \quad \text{(D.7)}$$

where K_1 is defined in Lemma 3, Appendix C, using Theorem 3.2 and Lemma 3 we get, for large enough x, with $M(x)=x^{1/4}$,

$$\mathbb{P}^\star\left(|\frac{v_n}{n}\sum_{i=1}^n w_p(\widehat{\theta}_n, Y_i)1\!\!1_{D_p}(Y_i)1\!\!1_{\|\theta-\widehat{\theta}\|\le M_0\sqrt{\log n/n}}|\ge x\right)\le$$
$$2\exp\left(-\frac{nx}{v_nC_n(x)}\right)+2\exp\left(-\frac{nx^2}{v_n^2\tau_n(x)^2}\right)+8\exp\left(-c^\star x^{1/2}\right)$$

with

$$\tau_n(x)^2=16C_1^2\sup_{\sqrt{n}\|\theta-\theta^\star\|\le M(x)\wedge M_0\sqrt{\log n}}E_{\mathbb{P}^\star}[w_p^2(\theta,Y_1)1\!\!1_{D_p}(Y_1)],$$
$$C_n(x)=4C_2C^\star c\left(W_{n,p,x}\right)C\left(W_{n,p,x}\right),$$

where $W_{n,p,x}$ is such that

$$\sup_{\sqrt{n}\|\theta-\theta^\star\|\le M(x)\wedge M_0\sqrt{\log n}} w_p(\theta,\cdot)1\!\!1_{D_p}\le W_{n,p,x}(\cdot).$$

For instance we may take

$$W_{n,p,x}\left(y\right)=\log p+C\frac{(|y|+A_p)M(x)}{\sqrt{n}b_p^2},$$

leading, by choosing $c(W_{n,p,x})=C\frac{M(x)}{\sqrt{n}b_p^2\log n}$, to

$$C_n(x)=C\frac{A_px^{1/2}}{\sqrt{n}b_p^2\log n}. \quad \text{(D.8)}$$

For any θ set

$$s_{p,\theta}(y)=\sum_{j=1}^{k^\star}\mu(j)f_p(y-m_j)\ \text{ and }\ s_\theta^\star(y)=\sum_{j=1}^{k^\star}\mu(j)f^\star(y-m_j).$$

We consider the following decomposition,

$$\log\left(\frac{s_{p,\theta}(y)}{s_{p,\theta^\star}(y)}\right)=\log\left(\frac{s_{p,\theta}(y)}{s^\star_\theta(y)}\right)+\log\left(\frac{s^\star_\theta(y)}{s^\star(y)}\right)+\log\left(\frac{s^\star(y)}{s_{p,\theta^\star}(y)}\right). \tag{D.9}$$

The first and third terms of (D.9) are treated similarly. (D.1) gives that for any θ, over D_p,

$$\left|\log\left(\frac{s_{p,\theta}(y)}{s^\star_\theta(y)}\right)\right|\lesssim R(y)b_p^\beta. \tag{D.10}$$

For the second term, since $f^\star\in H_{loc}(\beta,\gamma,\mathcal{P})$ with $\beta\geq 1/2$,

$$|\log f^\star(y-\widehat{m}_j)-\log f^\star(y-m^\star_j)|\leq L(y-m^\star_j)|\widehat{m}_j-m^\star_j|^{\beta\wedge 1}.$$

Moreover, if $y\in D_p$, and $\sqrt{n}\|\theta-\theta^\star\|\leq M(x)\wedge M_0\sqrt{\log n}$, then for large enough n,

$$L(y-m^\star_j)|\widehat{m}_j-m^\star_j|^{\beta\wedge 1}\leq 1$$

so that we have, for θ such that $\sqrt{n}\|\theta-\theta^\star\|\leq M(x)\wedge M_0\sqrt{\log n}$, over D_p, for large enough n,

$$\begin{aligned}\left|\log\left(\frac{s^\star_\theta(y)}{s^\star(y)}\right)\right|&\lesssim\sum_j\frac{|\mu_j-\mu^\star_j|}{\mu^\star_j}+\sum_j|m_j-m^\star_j|^{\beta\wedge 1}L(y-m^\star_j)\\&\lesssim\frac{M(x)}{\sqrt{n}}+(n^{-1/2}M(x))^{\beta\wedge 1}\sum_j L(y-m^\star_j).\end{aligned} \tag{D.11}$$

Thus, using the fact that $\beta\geq 1/2$, for large enough x,

$$\sup_{\sqrt{n}\|\theta-\theta^\star\|\leq M(x)\wedge M_0\sqrt{\log n}}E_{\mathbb{P}^\star}[w_p^2(\theta,Y_1)1\!\!1_{D_p}(Y_1)]=O(M(x)^2b_p^{2\beta}). \tag{D.12}$$

(D.8) and (D.12) give that, for all $\beta\geq 1/2$, for large enough x,

$$\frac{nx}{v_nC_n(x)}\gtrsim x^{1/2}n^{3/2}b_p^{2\beta+2}\gtrsim x^{1/2}(\log n)^{3(2\beta+2)/(2\beta+1)},\quad\frac{nx^2}{v_n^2\tau_n^2(x)}\gtrsim x^{1/2}n,$$

so that for large enough x,

$$\mathbb{P}^\star\left(\frac{v_n}{n}\sum_{i=1}^n w_p(\widehat{\theta}_n,Y_i)1\!\!1_{D_p}(Y_i)1\!\!1_{\|\theta-\widehat{\theta}\|\leq M_0\sqrt{\log n/n}}\geq x\right)\lesssim\exp\left(-Cx^{1/2}\right) \tag{D.13}$$

as soon as (D.6) and (D.7) hold for large enough x.

We now prove (D.6).

$$\begin{aligned}E_{\mathbb{P}^\star}[w_p(\theta,Y_1)1\!\!1_{D_p}(Y_1)]&=\int_{D_p}(s^\star(y)-s_{p,\theta^\star}(y))\log\left(\frac{s_{p,\theta}(y)}{s_{p,\theta^\star}(y)}\right)dy-K(s_{p,\theta^\star},s_{p,\theta})\\&\leq\int_{D_p}(s^\star(y)-s_{p,\theta^\star}(y))\log\left(\frac{s_{p,\theta}(y)}{s_{p,\theta^\star}(y)}\right)dy.\end{aligned}$$

Moreover, (D.1) and (D.10) give that

$$\begin{aligned}&\int_{D_p}|s^*(y)-s_{p,\theta^\star}(y)|\left|\log\left(\frac{s_{p,\theta}(y)}{s^\star_\theta(y)}\right)\right|dy\\&\qquad\lesssim b_p^\beta h(s^\star,s_{p,\theta^\star})\left(\int_{D_p}|R(y)|^2(s^\star(y)+s_{p,\theta^\star}(y))dy\right)^{1/2}\lesssim b_p^{2\beta}\end{aligned}$$

using (D.2). Also, (D.1) and (D.11) give that for θ such that $\sqrt{n}\|\theta-\theta^\star\| \le M(x) \wedge M_0\sqrt{\log n}$,

$$\int_{D_p} |s^\star(y) - s_{p,\theta^\star}(y)| \left|\log\left(\frac{s^\star_\theta(y)}{s^\star(y)}\right)\right| dy \lesssim b_p^{2\beta} M(x)^{\beta\wedge 1}$$

so that for $\beta \ge 1/2$, uniformly for θ such that $\sqrt{n}\|\theta - \theta^\star\| \le M(x) \wedge M_0\sqrt{\log n}$,

$$E_{\mathbb{P}^\star}[w_p(\theta, Y_1)1\!\!1_{D_p}(Y_1)] = O(M(x)b_p^{2\beta})$$

and (D.6) holds for large enough x.

To Prove (D.7) we use (D.9). We first control

$$\mathbb{E}_{\mathbb{P}^\star}\left(\sup_{\sqrt{n}\|\theta-\theta^\star\|\le M(x)\wedge M_0\sqrt{\log n}} \mathbf{G}_n(1\!\!1_{D_p}\log(s_{p,\theta}/s^\star_\theta))\right).$$

Using (D.10), we can bound on D_p,

$$\left|\log\left(\frac{s_{p,\theta}}{s^\star_\theta}(y)\right)\right| \lesssim |R(y)|b_p^\beta \lesssim (\log p)^{-1/2} \le 1$$

for n large enough, uniformly over $\sqrt{n}\|\theta-\theta^\star\| \le M(x)\wedge M_0\sqrt{\log n}$. Also, $\|f\|^2_{2,\beta} \lesssim \int f^2(1+\log^+|f|)(y)dy \lesssim \|f\|_2^2$, for any f in the form $\log(s_{p,\theta}/s^\star_\theta)$. We denote

$$\varphi_1(\sigma) = \int_0^\sigma \sqrt{\mathcal{H}(u, S_{n,p,1}(\sigma), \|.\|_2)}du$$

with

$$S_{n,p,1}(\sigma, x) = \{\log(s_{p,\theta}/s^\star_\theta), \sqrt{n}\|\theta - \theta^\star\| \le M(x) \wedge M_0\sqrt{\log n}, \|\log(s_{p,\theta}/s^\star_\theta)\|_2 \le \sigma\}.$$

Then for all $y \in D_p$, since $|y| \le A_p$, and for all $|m_j - m'_j| \le \eta$,

$$\begin{aligned} f_p(y - m'_j) &= \sum_{l=1}^p \pi_l \varphi_{b_p}(y - m_j - \alpha_l) e^{-\frac{(m_j - m'_j)^2}{2b_p^2} + \frac{(y - m_j - \alpha_l)(m_j - m'_j)}{b_p^2}} \\ &\le f_p(y - m_j) e^{\frac{(|y| + m_{k^\star} + A_p)\eta}{b_p^2}} \\ &\le f_p(y - m_j) e^{\frac{3A_p\eta}{b_p^2}} := f_U(y - m_j) \\ &\ge f_p(y - m_j) e^{-\frac{\eta^2}{2b_p^2} - \frac{3A_p\eta}{b_p^2}} := f_L(y - m_j) \end{aligned}$$

and

$$\begin{aligned} f^\star(y - m'_j) &\le f^\star(y - m_j) e^{\eta^{\beta\wedge 1}\sup_{|m - m_j| < \eta}|\tilde{\ell}(y - m)|} \\ &\ge f^\star(y - m_j) e^{-\eta^{\beta\wedge 1}\sup_{|m - m_j| < \eta}|\tilde{\ell}(y - m)|} \end{aligned} \tag{D.14}$$

where $\tilde{\ell}(y - m) = \ell_1(y - m)$ if $\beta > 1$ and $\tilde{\ell}(y - m) = L(y - m_j)$ if $\beta \le 1$, so that a bracket for $\log(s_{p,\theta'}/s^*\theta')1\!\!1_{D_p}$ is given on D_p by

$$\begin{aligned} U_{p,\theta} &:= \left(\frac{3A_p\eta}{b_p^2} + \eta^{\beta\wedge 1}\sup_{|m - m_j| < \eta}|\tilde{\ell}(y - m)|\right) + \log(1 + \eta\sum_{j=1}^{k^\star}\mu(j)^{-1}) \\ L_{p,\theta} &:= -\frac{3A_p\eta}{b_p^2} - \eta^{\beta\wedge 1}\sup_{|m - m_j| < \eta}|\tilde{\ell}(y - m)| + \log(1 - \eta\sum_{j=1}^{k^\star}\mu(j)^{-1}), \end{aligned}$$

Thus if $u > 0$ and $\eta \le \eta_0(u^{1/2}b_p^2/A_p \wedge u^{(\beta\wedge 1)/2})$ with $\eta_0 > 0$ small enough,

$$\int_{D_p} (U_{p,\theta} - L_{p,\theta})^2(y) s^\star(y) dy \le u,$$

so that

$$\varphi_1(\sigma) \lesssim \sigma\sqrt{\log^+(1/\sigma) + \log(nM(x))}.$$

Moreover for all $\|\theta - \theta^\star\| \leq M_0\sqrt{\log n}/\sqrt{n}$, (D.1) implies that

$$\| \log(s_{p,\theta'}/s^\star_{\theta'})\|_2^2 \leq b_p^{2\beta}C.$$

Therefore using Theorem 2 of Doukhan et al. (1995) and the fact that the chain is geometrically ergodic, we obtain that

$$\begin{aligned}\mathbb{E}_{\mathbb{P}^\star}\left(\sup_{\sqrt{n}\|\theta-\theta^\star\|\leq M(x)\wedge M_0\sqrt{\log n}} \mathbf{G}_n(1\!\!1_{D_p}\log(s_{p,\theta}/s^\star_\theta))\right) &\lesssim b_p^\beta(\log n + \log M(x)) \\ &\leq x/8,\end{aligned} \tag{D.15}$$

for $x \geq 1$ and large enough n. We now study

$$\mathbb{E}_{\mathbb{P}^\star}\left(\sup_{\sqrt{n}|\theta-\theta^\star\|\leq M(x)\wedge M_0\sqrt{\log n}} \mathbf{G}_n\left[1\!\!1_{D_p}\log\left(\frac{s^\star_\theta(y)}{s^\star(y)}\right)\right]\right).$$

Using (D.14), if $\sqrt{n}|\theta - \theta^\star\| \leq M(x) \wedge M_0\sqrt{\log n}$,

$$\left|\log\left(\frac{s^\star_\theta(y)}{s^\star(y)}\right)\right| \lesssim b_p^{-(\beta\wedge 1)}\sqrt{\log n}/\sqrt{n} = o(1),$$

so that

$$\begin{aligned}\left\|\log\left(\frac{s^\star_\theta(y)}{s^\star(y)}\right)\right\|^2_{2,\beta} &\lesssim \left\|\log\left(\frac{s^\star_\theta(y)}{s^\star(y)}\right)\right\|^2_2 \\ &\lesssim \max_j(\mu_j/\mu^\star_j - 1)^2 + \max_j \int s^\star(y)(\log f^\star(y-m_j) - \log f^\star(y-m^\star_j))^2 dy \\ &\lesssim (M(x)^2/n)^{\beta\wedge 1}.\end{aligned}$$

Hence using the same tricks as before and applying Theorem 2 of Doukhan et al. (1995) we obtain that for large enough n,

$$\mathbb{E}_{\mathbb{P}^\star}\left(\sup_{\sqrt{n}\|\theta-\theta^\star\|\leq M(x)\wedge M_0\sqrt{\log n}} \mathbf{G}_n(1\!\!1_{D_p}\log(s_{p,\theta}/s^\star_\theta))\right) \lesssim M(x)n^{-(\beta\wedge 1)/2}\sqrt{\log n} = o(x\sqrt{n}/v_n) \tag{D.16}$$

for all x and (D.7) is satisfied.
Finally, (D.13) holds, which, together with (D.5) ends the proof of Theorem 4.2.

References

Adamczak, R. and Bednorz, W. (2012). Exponential concentration inequalities for additive functionals of Markov chains. Technical report, University of Warsaw.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. pages 267–281. Second International Symposium on Information Theory.

Allman, E. S., Matias, C., and Rhodes, J. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37:3099–3132.

Azzaline, A. and Bowman, A. W. (1990). A look at some data in the Old Faithful geyser. *Applied Statistics*, 39:357–365.

Bonhomme, S., Jochman, K., and Robin, J. (2011). Nonparametric estimation of finite mixtures. Technical report, Department of Statistics.

Butucea, C. and Vandekerkhove, P. (2011). Semiparametric mixtures of symmetric distributions. Technical report.

Chambaz, A., Garivier, A., and Gassiat, E. (2009). A MDL approach to HMM with Poisson and Gaussian emissions. Application to order indentification. *Journal of Stat. Planning and Inf.*, 139:962–977.

Chambaz, A. and Rousseau, J. (2008). Bounds for Bayesian order identification with application to mixtures. *Ann. Statist.*, 36:938–962.

Clemencon, S., Garivier, A., and Tressou, J. (2009). Pseudo-regenerative block-bootstrap for hidden markov chains. In *Statistical Signal Processing, 2009. SSP '09 IEEE.*

Doukhan, P., Massart, P., and Rio, E. (1994). The functional central limit theorem for strongly mixing processes. *Annales de l'I.H.P.*, 30:63–82.

Doukhan, P., Massart, P., and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Annales de l'I.H.P.*, 31:393–427.

Gassiat, E. and van Handel, R. (to appear). Consistent order estimation and minimal penalties. *IEEE Trans. Info. Theory.* http://arxiv.org/abs/1002.1280.

Hall, P. and Zhou, X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.*, 31:201–224.

Hunter, D. R., Wang, S., and Hettmanspeger, T. P. (2004). Inference for mixtures of symmetric distributions. *Ann. Statist.*, 35:224–251.

Ishwaran, H., James, L., and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. American Statist. Assoc.*, 96(456):1316–1332.

Kasahara, H. and Shimotsu, K. (2007). Nonparametric identification and estimation of multivariate mixtures. *preprint*, pages 1–26.

Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). Adaptive bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, pages 1225–1257.

L. Bordes, S. M. and Vandekerkhove, P. (2006). Semiparametric estimation of a two components mixture model. *Annals of Statistics*, 34:1204–1232.

Lambert, M. F., Whiting, J. P., and Metcalfe, A. V. (2003). A non-parametric hidden Markov model for climate state identification. *Hydrology and earth system sciences*, 7:652–667.

MacLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* John Wiley, New York.

Marin, J.-M., Mengersen, K., and Robert, C. (2005). Bayesian modelling and inference on mixtures of distributions. In Rao, C. and Dey, D., editors, *Handbook of Statistics*, volume 25. Springer-Verlag, New York.

Massart, P. (2007). *Concentration Inequalities and Model Selection : Ecole d'Et de Probabilits de Saint-Flour XXXIII - 2003.* Berlin ; Heidelberg (DEU) ; New York : Springer.

Maugis, C. and Michel, B. (2011). Data-driven penalty calibration: A case study for gaussian mixture model selection. *ESAIM : P&S*, 15:320–339.

Maugis-Rabusseau, C. and Michel, B. (2012). Adaptive density estimation using finite gaussian mixtures. *ESAIM : P&S*, page to appear.

Moreno, C. (1973). The zeros of exponential polynomials (i). *Compositio Mathematica*, 26:69–78.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Society Series B*, 59:731–792.

Rio, E. (2000). Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus Acad. Sciences Paris*, 330:905–908.

Stein, E. M. and Shakarchi, R. (2003). *Complex Analysis*. Princeton University Press, Princeton.

Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2011). Bayesian nonparametric hidden Markov models with applications in genomics. *J. Royal Statist. Society Series B*, 73:1–21.